eKNOW 2011

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

February 23-28, 2011 - Gosier

Guadeloupe, France

eKNOW 2011 Editors

Borka Jerman-Blazic, Jozef Stefan Institute, Slovenia

Dumitru Dan Burdescu, University of Craiova, Romania

eKNOW 2011

Foreword

The Third International Conference on Information, Process, and Knowledge Management [eKNOW 2011], held between February 23-28, 2011 in Gosier, Guadeloupe, France, continued the series of events dealing with the management of information, process and knowledge in today's complex environments.

The variety of systems and applications and the heterogeneous nature of information and knowledge representation require special technologies to capture, manage, store, preserve, interpret and deliver the content and documents related to a particular target.

Progress in cognitive science, knowledge acquisition, representation, and processing helped to deal with imprecise, uncertain or incomplete information. Management of geographical and temporal information becomes a challenge, in terms of volume, speed, semantic, decision, and delivery.

Information technologies allow optimization in searching and interpreting data, yet special constraints imposed by the digital society require on-demand, ethics, and legal aspects, as well as user privacy and safety.

Nowadays, there is notable progress in designing and deploying information and organizational management systems, expert systems, tutoring systems, decision support systems, and in general, industrial systems.

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

We take here the opportunity to warmly thank all the members of the eKNOW 2011 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 eKNOW 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the eKNOW 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that eKNOW 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of information, process and knowledge management.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the beautiful surroundings of Gosier, Guadeloupe, France.

eKNOW 2011 Chairs

Susan Gauch, University of Arkansas, USA
Roy Oberhauser, Aalen University, Germany
Borka Jerman-Blazic, Jozef Stefan Institute, Slovenia
Jeff Riley, Hewlett-Packard Australia, Australia
Gil ad Ariely, Lauder School of Government / Interdisciplinary Center Herzliya (IDC), Israel
Christian Bartsch, Research Center for Information Technology (FZI) - Karlsruhe, Germany
Pierre-N. Robillard, Ecole Polytechnique de Montréal, Canada
Ernesto Exposito, INSA/DGEI & LAAS/CNRS - Toulouse, France

eKNOW 2011

Committee

eKNOW Advisory Committee

Susan Gauch, University of Arkansas, USA
Roy Oberhauser, Aalen University, Germany
Borka Jerman-Blazic, Jozef Stefan Institute, Slovenia
Jeff Riley, Hewlett-Packard Australia, Australia
Gil ad Ariely, Lauder School of Government / Interdisciplinary Center Herzliya (IDC), Israel
Christian Bartsch, Research Center for Information Technology (FZI) - Karlsruhe, Germany
Pierre-N. Robillard, Ecole Polytechnique de Montréal, Canada
Ernesto Exposito, INSA/DGEI & LAAS/CNRS - Toulouse, France

eKNOW 2011 Technical Program Committee

Werner Aigner, Institute for Application Oriented Knowledge Processing - FAW / University of Linz, Austria
Gil ad Ariely, Lauder School of Government / Interdisciplinary Center Herzliya (IDC), Israel
Ezendu Ariwa, London Metropolitan University, United Kingdom
Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong
Christian Bartsch, Research Center for Information Technology (FZI) - Karlsruhe, Germany
Ladjel Bellatreche, LISI- ENSMA/ Poitiers University, France
Peter Bellström, Karlstad University, Sweden
Jorge Bernardino, ISEC - Institute Polytechnic of Coimbra, Portugal
Carsten Brockmann, Universität Potsdam, Germany
Sabine Bruaux, Picardie Jules Verne University, France
Martine Cadot, University of Nancy1, France
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Expedito Carlos Lopes, Federal University of Campina Grande, Brazil
Petre Dini, Cisco Systems, Inc., USA / Concordia University, Canada
Ernesto Exposito, INSA/DGEI & LAAS/CNRS) - Toulouse, France
Susan Gauch, University of Arkansas, USA
Olivier Gendreau, École Polytechnique de Montréal, Canada
Conceição Granja, Siemens S.A. / Universidade do Porto, Portugal
Manfred Grauer, University of Siege, Germany
Pierre Hadaya, ESG UQAM, Canada
Céline Hudelot, Ecole Centrale Paris, France
Khaled Khelif, EADS- Val de Reuil, France
Marite Kirikova, Riga Technical University, Latvia
Agnes Koschmider, KIT, Germany
Andrew Kusiak, The University of Iowa, USA
Hiep Luong, University of Arkansas, USA
Dirk Malzahn, OrgaTech GmbH, Germany
Marco Mevius, HTWG Konstanz, Germany
Roy Oberhauser, Aalen University, Germany

Daniel O'Leary, University of Southern California, USA
Zinayida Petrushyna, RWTH - Aachen, Germany
Jeff Riley, Hewlett-Packard Australia, Australia
Kenji Saito, Keio University, Japan
Erwin Schaumlechner, Technology Center Tiscover AG, Austria
Tim Schlüter, Heinrich Heine University - Düsseldorf, Germany
Pnina Soffer, University of Haifa, Israel
Lubomir Stanchev, Indiana University - Purdue University Fort Wayne, USA
Carlo Tasso , Università di Udine, Italy
Lars Taxén, Linköpings Universitet-Tullinge, Sweden
Andrea Valente, Aalborg University - Esbjerg, Denmark
Jan Martijn van der Werf, Technische Universiteit Eindhoven, The Netherlands
Aurora Vizcaino Barcelo, University of Castilla-La Mancha, Spain
Meng Yu, Virginia Commonwealth University, USA

Copyright Information

Table of Contents

Abstraction of Informed Virtual Geographic Environments for the Modeling of Large-Scale and Complex Geographic Environments

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

***Abstract*—In this paper, we propose a semantically-informed and geometrically-accurate virtual geographic environment method which allows to use Geographic Information System (GIS) data to automatically built an Informed Virtual Geographic Environment (IVGE). Besides, we propose an abstraction process which uses geometric, toplogic, and semantic characteristics of geographic features in order to build a knowledge-based description of the IVGE relying on a hierarchical graph-based structure. Our IVGE model enables the support of large-scale and complex geographic environments modeling for Situated Multi-Agent Systems (SMAS) in which agents are situated and with which they interact.**

***Keywords*-Informed Virtual Geographic Environments; Environmental Abstraction; Knowledge Representation.**

I. Introduction

During the last decade, the Multi-Agent Geo-Simulation (MAGS) approach has attracted a growing interest from researchers and practitioners to simulate phenomena in a variety of domains including traffic simulation, crowd simulation, urban dynamics, and changes of land use and cover, to name a few [1]. Such approaches are used to study phenomena (i.e., car traffic, mobile robots, sensor deployment, crowd behaviors, etc.) involving a large number of simulated actors (implemented as software agents) of various kinds evolving in, and interacting with, an explicit description of the geographic environment called Virtual Geographic Environment (VGE).

A critical step towards the development of a MAGS is the creation of a VGE, using appropriate representations of the geographic space and of the objects contained in it, in order to efficiently support the agents' situated reasoning. Since a geographic environment may be complex and large scale, the creation of a VGE is difficult and needs large quantities of geometrical data originating from the environment characteristics (terrain elevation, location of objects and agents, etc.) as well as semantic information that qualifies space (building, road, park, etc.).

In order to yield realistic MAGSs, a VGE must precisely represent the geometrical information which corresponds to geographic features. It must also integrate several semantic notions about various geographic features. To this end, we propose to enrich the VGE data structure with semantic information that is associated with the geographic features. Moreover, we propose to abstract this semantically-enriched and geometrically-precise VGE description in order to enable large-scale and complex geographic environments modeling.

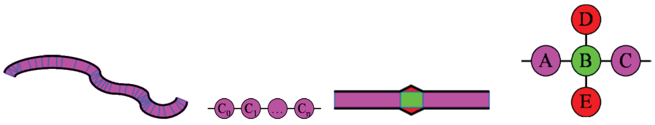
In this paper, we present a novel approach that addresses these challenges toward the creation of such a semantically-enriched and geometrically-accurate VGE, which we call an *Informed VGE* (IVGE). We also detail our abstraction technique to support large-scale and complex geographic environments. The rest of the paper is organized as follows: Section II provides an overview of related works. Section III introduces our IVGE computation model. Section IV presents the proposed abstraction approach which is composed of the three processes; (1) geometric abstraction; (2) topologic abstraction; and (3) semantic abstraction. Section V discusses the proposed abstraction approach. Finally, Section VI concludes and presents the future perspectives of this work.

II. Related Works

Virtual environments and spatial representations have been used in several application domains. For example, Thalmann *et al.* proposed a virtual scene for virtual humans representing a part of a city for graphic animation purposes [3]. Donikian *et al.* proposed a modelling system which is able to produce a multi-level data-base of virtual urban environments devoted to driving simulations [15]. More recently, Shao *et al.* proposed a virtual environment representing the New York City's Pennsylvania Train Station populated by autonomous virtual pedestrians in order to simulate the movement of people [13]. Paris *et al.* also proposed a virtual environment representing a train station populated by autonomous virtual passengers, in order to characterize the levels of services inside exchange areas [12]. However, since the focus of these approaches is computer animation and virtual reality, the virtual environment usually plays the role of a simple background scene in which agents mainly deal with geometric characteristics. Indeed, the description of the virtual environment is often limited to the geometric level, though it should also contain topological and semantic information for other types of applications using advanced agent-based simulations. Current virtual environment models do not support large-scale and complex geographic environments and fail to capture real world physical environments' characteristics. When dealing with large-scale and complex geographic environments, the spatial subdivision which can be either exact or approximate produces a large number of cells [7]. The topologic approach allows representation of such a spatial subdivision using a graph structure and to take advantage of efficient algorithms provided by the graph theory [12]. However, the graph size may still remain large when dealing with geographic environments with dense geographic features [7]. Moreover, geographic features with curved geometries (*Figure 1*) produce a large number of triangles since they are initially represented by a large number of segments.

An *environment abstraction* is a process used to better organize the information obtained at the time of spatial subdivision of the geographic environment. The unification process is addressed principally in two ways: (1) a *pure topological* [8] unification which associates the subdivision cells according to their number of connexions; (2) a more *conceptual* unification which introduces a semantical definition of the environment, like with the *IHT-graph structure* [15]. Lamarche and Donikian proposed a topologic abstraction approach which assigns to each node of the graph resulting from the space decomposition a *topological qualification* according to the number of connected edges given by its arity [8]. The topologic abstraction algorithm aims to generate an abstraction tree by merging interconnected cells while trying to preserve topological properties [8]. When merging several cells into a single one, the composition of cells is stored in a graph structure in order to generate the abstraction tree. The topologic abstraction proposed by Lamarche and Donikian relies on the topological properties of the cells and reduces the size of the graph that represents the space subdivision [8]. However, the topological characteristics are not sufficient to abstract a virtual environment when dealing with a large-scale and complex environment involving areas with various qualifications (buildings, roads, parks, sidewalks, etc.).

Not much research has been done on semantic integration in the description of a virtual environment. The *Computer Animation* and *Behavioral Animation* research fields provide a few attempts to integrate the semantic information in order to assist agents interacting with their environments. Semantic information has been used for different purposes, including the simulation of inhabited cities [3], computer animation [6], and simulation of virtual humans [4]. Farenc has first used the notion of *Informed Environments* [3]. She defined informed environments as a database which represents urban environments with semantic information representing urban knowledge [3]. An informed environment is thus characterized as a place where information (semantic and geometrical) is dense, and can be structured and organized using rules [3]. Building an informed environment as presented by Farenc consists of adding a semantic layer onto a core corresponding to a classical scene (a set of graphical objects) modeled using graphical software for computer animations purposes [3].



(a) Curved geometries. (b) Alignment anomalies.

Figure 1: Cells resulting from curved geometries (a) and alignment anomalies (b) [12].

Despite the multiple designs and implementations of virtual environments frameworks and systems, the creation of geometrically-accurate and semantically-enriched geographic content is still an open issue. Indeed, research has focused almost exclusively on the geometric and topologic characteristics of the virtual geographic environment. However, the structure of the virtual environment description, the optimization of this description to support large-scale and complex geographic environments, the meaning of the geographic features contained in the environment as well as the ways to interact with them have received less attention.

III. Computation of IVGE

In this section, we briefly present our automated approach to compute the IVGE data using vector GIS data. This approach is based on four stages: *input data selection*, *spatial decomposition*, *maps unification*, and finally the generation of the *informed topologic graph* [10].

GIS Input Data Selection: The first step of our approach consists of selecting the different vector data sets which are used to build the IVGE. The input data can be organized into two categories. First, *elevation layers* contain geographical marks indicating absolute terrain elevations. Second, *semantic layers* are used to qualify various types of data in space. Each layer indicates the physical or virtual limits of a given set of features with identical semantics in the geographic environment, such as roads or buildings.

Spatial Decomposition: The second step consists of obtaining an exact spatial decomposition of the input data into cells. First, an elevation map is computed using the

Constrained Delaunay Triangulation (CDT) technique. All the elevation points of the layers are injected into a 2D triangulation, the elevation being considered as an attribute of each node. Second, a merged semantics map is computed, corresponding to a constrained triangulation of the semantic layers. Indeed, each segment of a semantic layer is injected as a constraint which keeps track of the original semantic data by using an additional attribute for each semantic layer.

Map Unification: The third step to obtain our IVGE consists of unifying the two maps previously obtained. This phase can be depicted as mapping the $2D$ merged semantic map onto the $2.5D$ elevation map in order to obtain the final $2.5D$ elevated merged semantics map. First, preprocessing is carried out on the merged semantics map in order to preserve the elevation precision inside the unified map. Indeed, all the points of the elevation map are injected into the merged semantics triangulation, creating new triangles. Then, a second process elevates the merged semantics map.

Informed Topologic Graph: The resulting unified map now contains all the semantic information of the input layers, along with the elevation information. This map can be used as an *Informed Topologic Graph* (ITG), where each node corresponds to the map's triangles, and each arc corresponds to the adjacency relations between these triangles. Then, common graph algorithms can be applied to this topological graph, and graph traversal algorithms in particular.

IV. Abstraction of IVGE

In this Section, we describe the abstraction process which optimizes the description of the IVGE. Sub-section IV-A presents the first enhancement which is related to the qualification of terrain. We propose a novel approach of information extrapolation using a one-time spatial reasoning process based on a geometric abstraction. This approach can be used to fix input elevation errors, as well as to create new qualitative data relative to elevation variations. These data are stored as additional semantics bound to the graph nodes, which can subsequently be used for spatial reasoning. Sub-section IV-B introduces the second enhancement which optimizes the size of the informed graph structure using a topological abstraction process. This process aims at building an hierarchical topologic graph structure in order to deal with large-scale virtual geographic environments. Sub-section IV-C details the third enhancement technique which propagates qualitative input information from the arcs of the graph to the nodes, which allows deduction of the internal parts of features such as buildings or roads in addition to their boundaries. Moreover, this technique uses Conceptual Graphs (CG) [14], a standard formalism for the representation of semantic information. Figure 2 illustrates the abstracted IVGE generation model.

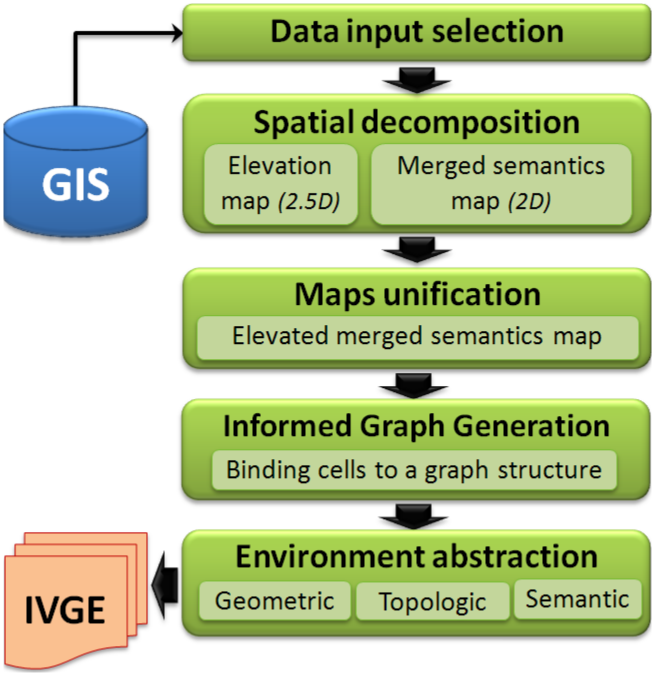


Figure 2: The IVGE global architecture of IVGE generation including the environment abstraction process.

A. Geometric abstraction

Spatial decomposition subdivides the environment into convex cells. Such cells encapsulate various quantitative geometric data which are suitable for precise computations. Since geographic environments are seldom flat, it is important to consider the terrain's elevation and shape. While elevation data are stored in a quantitative way which is suitable for exact calculations, spatial reasoning often needs to manipulate qualitative information. Indeed, when considering a slope, it is obviously simpler and faster to qualify it using an attribute with ordinal values such as *gentle* and *steep* rather than using numerical values. However, when dealing with large scale geographic environments, handling the terrain's elevation, including its light variations, may be a complex task. To this end, we propose an abstraction process that uses geometric data to extract the average terrain's elevation information from spatial areas. The objectives of this *Geometric Abstraction* are threefold. First, it aims to reduce the amount of data used to describe the environment. Second, it helps for the detection of anomalies, deviations, and aberrations in elevation data. Third, the geometric abstraction enhances the environment description by integrating qualitative information characterizing the terrain shape. In this section, we first present the algorithm which computes the geometric abstraction. Then, we describe two processes which use the geometric abstraction, namely *Filtering elevation anomalies* and *Extracting elevation semantics*.

1) Geometric Abstraction Algorithm: As presented in the previous chapter, the geographic environment is subdivided into cells of different shapes and sizes. The algorithm takes advantage of the graph structure obtained from the IVGE extraction process. A *cell* corresponds to a node in the topological graph. A node represents a triangle generated by the *CDT* spatial decomposition technique. A cell is characterized by its boundaries, its neighboring cells, its surface as well as its normal vector which is a vector perpendicular to its plane.

Now we introduce the notion of a group, which is a collection of adjacent cells. The grouping strategy is based on a coplanarity criterion which is assessed by computing the difference between the *normal vectors* of two neighboring cells or groups of cells. Since a group is basically composed of adjacent cells it is obvious to characterize a group by its boundaries, its neighboring groups, its surface, as well as its normal vector. However, the normal vector of a group must rely on an interpretation of the normal vectors of its composing cells. In order to compute the normal vector of a group, we adopt the *area-weight normal vector* [2], which takes into account the unit normal vectors of its composing cells as well as their respective surfaces. Let S_c denote the surface area of a cell c and $\overrightarrow{N_c}$ be its unit normal vector. The area-weight normal vector $\overrightarrow{N_G}$ of a group G is computed as follows:

$$\overrightarrow{N_G} = \sum_{c \in G} \left(S_c \ . \ \overrightarrow{N_c}\right) / \sum_{c \in G} S_c \qquad (1)$$

The geometric abstraction algorithm uses two input parameters: 1) a set of *starting cells* which act as access points to the graph structure, and 2) a Δ parameter which corresponds to the maximal allowed difference between cells' gradients. Two adjacent cells are considered coplanar, and hence grouped, when the angle between their normal vectors is lesser than Δ. The recursive geometric abstraction algorithm is composed of five steps:

1) For each cell c of the *starting cells*, create a new group G and do step 2.
2) For each neighbouring group or cell n of G, if the neighbour has already been processed, do step 3, else do step 5.
3) If angle $\left(\overrightarrow{N_G}, \overrightarrow{N_n}\right) \leq \Delta$ then do step 4. Otherwise do step 5.
4) Merge n in group G, then evaluate $\overrightarrow{N_G}$ using equation (1). Do step 2 again for G.
5) If n is an unprocessed cell, create a new group G with n and do step 2.

The algorithm starts by visiting all the cells of the virtual environment. For each visited cell *crt_cell*, a new group *crt_grp* is created and the cell is registered as a member (line 2). The area-weighted normal vector of *crt_grp* is computed using equation (1). Besides, the algorithm tests the coplanarity of *crt_cell* with its neighbouring cells (*nxt_cell* belonging to *nxt_grp*) using equation (1) and to decide whether to include these neighbours in the group *crt_grp* to which *crt_cell* belongs. Next, the algorithm explores and processes the neighbouring cells of *crt_cell*. For each neighbour, if it is visited for the first time, a new group is created and the neighbour cell is registered as its first member.

Afterwards, the algorithm computes the angles resulting from the merging of *crt_grp* and *nxt_grp* groups. The area-weighted normal vector resulting from the integration of *crt_grp*'s elements in the *nxt_grp* group is computed. The algorithm goes on by computing the angle between the new (after the merge) and the previous (before the merge) area-weighted normal vectors. The angle is given by the scalar product of the two normalised vectors $\overrightarrow{N}_{crt\_grp}$ and $\overrightarrow{N}_{nxt\_grp}$. If this angle respects the input parameter Δ (line 9), then merging is performed (line 10).

In the proposed algorithm, the geometric abstraction produces coherent groups whose cells are coplanar and with respect to the Δ threshold. The geometric abstraction process abstracts a higher-order topologic graph and produces a new graph with fewer nodes which helps to enhance performance of spatial reasoning mechanisms.

The analysis of the resulting groups helps to identify anomalies in elevation data. Such anomalies need to be fixed in order to build a realistic virtual geographic environment. Furthermore, the average terrain slope which characterizes each group is a quantitative datum described using area-weighted normal vectors. Such quantitative data are too precise to be used by qualitative spatial reasoning. Hence, a qualification process would greatly simplify spatial reasoning mechanisms. Thus the geometric abstraction can improve IVGE by filtering the elevation anomalies, qualifying the terrain slope using semantics and integrating such semantics in the description of the geographic environment.

2) Filtering elevation anomalies: Analysis of the geometric abstraction may reveal an isolated group which is totally surrounded by another single coherent group. These groups are characterised by a large difference between their respective area-weighted normal vectors. Such isolated groups are often characterised by their small surface areas and can usually be considered as anomalies, deviations, or aberrations in the initial elevation data. MAGS users may verify if such groups correspond to real pits or depressions, or substantial mounds or heaps on the landscape. The geometric abstraction process helps to identify them and can help to automatically filter such anomalies using a two phase process. First, isolated groups are identified (*Figure 3(a)*). The identification of isolated groups is based on two key parameters: 1) the ratio between the surface areas of the surrounded and surrounding groups, and 2) the difference between the area-weighted normal vectors of the surrounded and surrounding groups. Second, these isolated groups are adjusted to the average level of elevation of the surrounding ones (*Figure 3(b)*). The lowest and the highest elevations (*low_elev*, *high_elev*) of the surrounding group (*surrounding_grp*) are computed. Then, the elevation of all the vertices of the isolated group (*isolated_grp*) are adjusted using the average between *lowest_elevation* and *highest_elevation*. As a consequence, we obtain more coherent groups in which anomalies of elevation data are corrected.

3) Qualification of terrain shape: The geometric abstraction algorithm computes quantitative geometric data which precisely describe the terrain However, handling and exploiting quantitative data is a complex task as the range of values may be too large and calculations or analysis methods may be too costly. Therefore, we propose to interpret the quantitative data representing the terrain shape by qualifying the terrain characteristics. Semantic labels, which are called *the shape semantics*, are associated to quantitative intervals of values that represent the terrain's shape. In order to obtain the shape's semantics we propose a two-step process taking advantage of the geometric abstraction: 1) calculation of the inclination, or the angle α between the weighted normal vector $\vec{N}_g$ of a group *grp* and the horizontal plane; and 2) assigning to each discrete value a semantic category which qualifies it. The discretisation process can be done in two ways: a *customised* and an *automated* approach.

The *customised approach* requires that the user provides a complete specification of the discretisation to qualify the range of slopes. Indeed, the user needs to specify a list of inclination intervals as well as their associated semantic labels. The algorithm iterates over the groups obtained by the geometric abstraction. For each group *grp*, it calculates the inclination value *I*. Then, this process checks the interval bounds and determines in which one the inclination value *I* falls. Finally, the customised discretisation extracts the semantic shape label from the selected slope interval and assigns it to the group *grp*. For example, let us consider the following inclination interval and the associated semantic label : $\{([10,\ 20],\ gentle\ slope), ([20,\ 25],\ steep\ slope)\}$. Such a customised specification associates the semantic label "*gentle slope*" to inclination values included in the interval $[10,\ 20]$ and the semantic label "*steep slope*" to inclination values included in the interval $[20,\ 25]$.

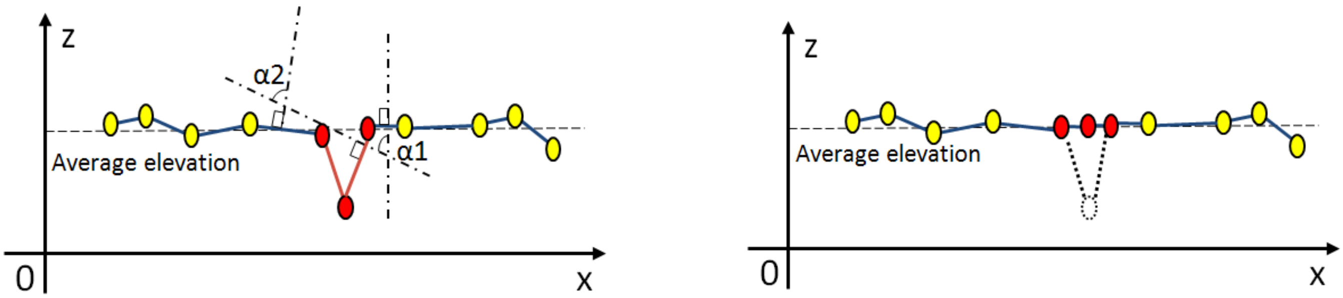
The *automated approach* only relies on a list of semantic shape labels representing the slope qualifications. Let N be the number of elements of this list, and T be the total number of groups obtained by the geometric abstraction algorithm. First, the automated discretisation orders groups based on their terrain inclination. Then, it iterates over the ordered groups and associates a uniform number of groups, T/N, to each semantic label from the *semantic set*, each T/N processed groups. For example, let us consider the following semantic slope labels: $\{gentle, medium, steep\}$, and an ordered set S of groups denoted as follows: $S = \{gr_i | i \in \{1, 2, .., 6\}\}$ with the following respective slope values: $\{5,\ 10,\ 15,\ 20,\ 25,\ 30\}$. For every 2 groups (as $T = 6$ and $N = 3$, $\frac{T}{N} = 2$), the automated discretisation assigns a new semantic slope label: $\{gentle, gentle, medium, medium, steep, steep\}$.

Let us compare these two discretisation approaches. On the one hand, the *custumized discretisation process* allows one to freely specify the qualification of the slopes, choosing ranges that match the problem domain. However, qualifications resulting from such a flexible approach deeply rely on the correctness of the interval bounds' values. Therefore, the customised discretisation method requires to have a good knowledge of the terrain characteristics in order to guarantee a valid specification of inclination intervals. On the other hand, the *automated discretisation process* is also able to qualify slopes without the need to specify interval bounds. This method also guarantees that all the specified semantic attributes will be assigned to the groups without a prior knowledge of the environment characteristics. However, the resulting intervals may have no relation to the problem domain.

4) Improving the geometric abstraction: Thanks to the extraction of slope semantics, terrain shape is qualified using semantic attributes and associated with groups and with their cells. Because of the nature of the classification intervals, adjacent groups with different area-weighted normal vectors may obtain the same semantic slope label. In order to improve the results provided by the geometric abstraction, we propose a process that merges adjacent groups which share the same semantic slope. This process starts by iterating over groups. Every time it finds a set of adjacent groups sharing an identical semantic slope, it creates a new group. Next, cells composing the adjacent groups are registered as members of the new group. Finally, the area-weighted normal vector is computed for the new group. Hence, this process guarantees that every group is only surrounded by groups which have different semantic slopes.

B. Topological abstraction

In Section III, we presented our work on the generation of informed virtual geographic environments using an exact spatial decomposition scheme which subdivides the environment into convex cells organized in a topological graph structure. However, inside large scale and complex geographic environments (such as a city for example), such topological graphs can become very large. The size of such a topological graph has a direct effect on paths'



(a) Identifying elevation anomalies. Two isolated groups (in red) and angles (α1 and α2) resulting from the difference between the area-weighted normal vectors

(b) Fixing elevation anomalies. Nodes in isolated groups are adjusted to the average elevation level.

Figure 3: Profile section of anomalous *Isolated Groups* (red colour) adjusted to the average elevation of the surrounding ones (yellow colour).

computation time for path-finding. In order to optimise the performance of path computation, we need to reduce the size of the topological graph representing the IVGE. The aim of the topological abstraction is to provide a compact representation of the topological graph that is suitable for situated reasoning and enables fast path planning. However, in contrast to the geometric abstraction which only enhances the description of the IVGE with terrain semantics, the topological abstraction extends the topological graph with new layers. In each layer (except for the initial layer which is called level 0), a node corresponds to a single or a group of nodes in the immediate lower level (*Figure 4*). The topological abstraction simplifies the IVGE description by combining cells (triangles) in order to obtain convex groups of cells. Such a hierarchical structure evolves the concept of *Hierarchical Topologic Graph* in which cells are fused into groups and edges are abstracted in boundaries. To do so, convex hulls are computed for every node of the topological graph. Then, the coverage ratio of the convex hull is evaluated as the surface of the hull divided by the actual surface of the node. The topological abstraction finally performs groupings of a set of connected nodes if and only if the group ratio is equal or close to one depending on the problem domain. Let C be the convexity rate and $CH(gr)$ be the convex hull of the polygon corresponding to gr. C is computed as follows:

$$C(gr) = \frac{Surface(gr)}{Surface(CH(gr))} \quad and \quad 0 < C(gr) \leq 1 \qquad (2)$$

The convex property of each group's hull needs to be preserved after the topological abstraction. This ensures that an entity can move freely inside a given cell (or group of cells), and that there exists a straight path linking edges belonging to the same cell (or group of cells).

Figure 5 illustrates an example of the topological abstraction process and the way it reduces the number of cells representing the environment. In *Figure 5(a)*, we present the initial vector format GIS data of a complex building.

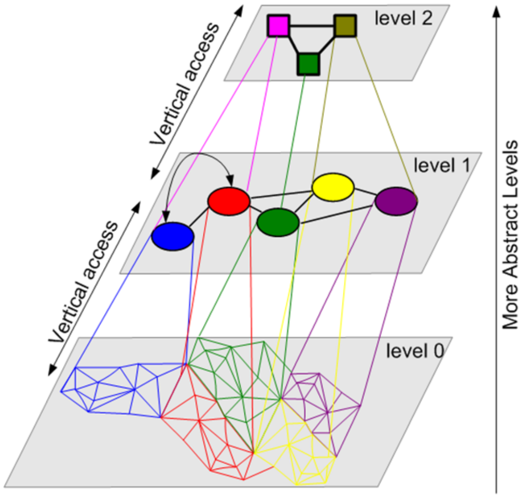


Figure 4: The topological graph extraction from space decomposition and extension into different levels using the topological abstraction.

Figure 5(b) depicts the initial exact spatial decomposition which yields 63 triangular cells. *Figure 5(c)* presents 28 convex polygons generated by the topological abstraction algorithm. The *abstraction rate* of the number of cells representing the environment is around 55%. This rate is computed using the ratio of initial number of cells produced by the space decomposition techniques (63) by the number of convex polygons (28) obtained using the topologic abstraction technique with a convexity rate equal to 1

To conclude, we described in this section a topologic abstraction process in order to enhance the performance of the exploration of the IVGE's description. This process aims to simplify large informed graphs corresponding to large-scale and complex geographic environments. Our topologic abstraction approach reduces the number of convex cells by overlaying the informed graph with a topologically abstracted graph. The resulting IVGE is hence based on a hierarchical graph whose lowest level corresponds to the informed graph initially produced by the spatial decomposition. In the following section, we show how we use a well-known knowledge representation formalism to represent the semantic information in order to further enhance the IVGE description with respect to agents' and the environment's characteristics.

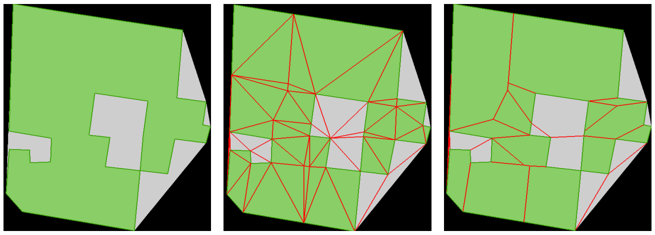

(a) (b) (c)

Figure 5: Illustration of the topological abstraction process with a strict convex property (*C(gr)* = 1); (a) the GIS data of a complex building; (b) the exact space decomposition using CDT techniques (63 triangular cells) ; (c) the topological abstraction (28 convex polygons)

C. Semantic Abstraction

Two kinds of information can be stored in the description of an IVGE. Quantitative data are stored as numerical values which are generally used to depict geometric properties (like a path's width of *2* meters) or statistical values (like a density of *2.5* persons per square meter). Qualitative data are introduced as identifiers which can range from a word with a given semantics, called a *label*, to a reference to an external database or to a specific knowledge representation. Such semantic information can be used to qualify an area (like a *road* or a *building*) or to interpret a quantitative value (like a *narrow* passage or a *crowded* place). An advantage of interpreting quantitative data is to reduce a potentially infinite set of inputs to a discrete set of values, which is particularly useful to condense information in successive abstraction levels to be used for reasoning purposes. Furthermore, the semantic information enhances the description of the IVGE, which in turn extends the agents' knowledge about their environment. However, the integration of the semantic information raises the issue of its representation. Therefore, we need a standard formalism that allows for precisely representing the semantic information which qualifies space and which is computationally tractable in order to be used by spatial reasoning algorithms used by agents.

Several knowledge representation techniques can be used to structure semantic information and to represent knowledge in general such as *frames* [11], *rules* [9] (also called *If-Then* rules), *tagging* [16], and *semantic networks* [14], which have originated from theories of human information processing. Since knowledge is used to achieve intelligent behavior, the fundamental goal of knowledge representation is to represent knowledge in a manner that facilitates inferencing (i.e., drawing conclusions) from knowledge. In order to select a knowledge representation (and a knowledge representation system to logically interpret sentences in order to derive inferences from them), we have to consider the expressivity of the knowledge representation. The more expressive a knowledge representation technique is, the easier (and more compact) we can describe and qualify geographic features which characterise IVGE. Various artificial languages and notations have been proposed to represent knowledge. They are typically based on logic and mathematics, and can be easily parsed for machine processing. However, Sowas's *Conceptual Graphs* [14] are widely considered an advanced standard logical notation for logic based on existential graphs proposed by Charles Sanders Peirce and on semantic networks.

Syntactically, a conceptual graph is a network of concept nodes linked by relation nodes. Concept nodes are represented by the notation *[Concept Type: Concept instance]* and relation nodes by *(Relationship-Name)*. A concept instance can be either a value, a set of values or even a CG. The formalism can be represented in either graphical or character-based notations. In the graphical notation, concepts are represented by rectangles, relations by circles and the links between concept nodes and relation nodes by arrows. The most abstract concept type is called the *universal type* (or simply *Universal*) denoted by the symbol ⊥.

A MAGS usually involves a large number of situated agents of different types (human, animal, static, mobile, etc.) performing various actions (moving, perceiving, etc.) in virtual geographic spaces of various extents. Using CGs greatly simplifies the representation of complex situated interactions occurring at different locations and involving various agents of different types. In order to create models for MAGS we consider three fundamental abstract concepts: 1) *agents*; 2) *actions*; and 3) *locations*.

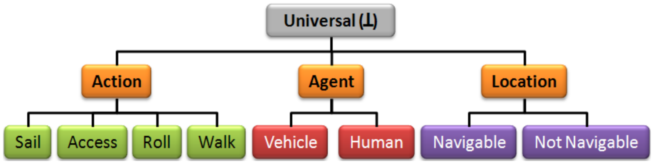


Figure 6: Illustration of the *action*, *agent* and *location* concepts using a concept type lattice.

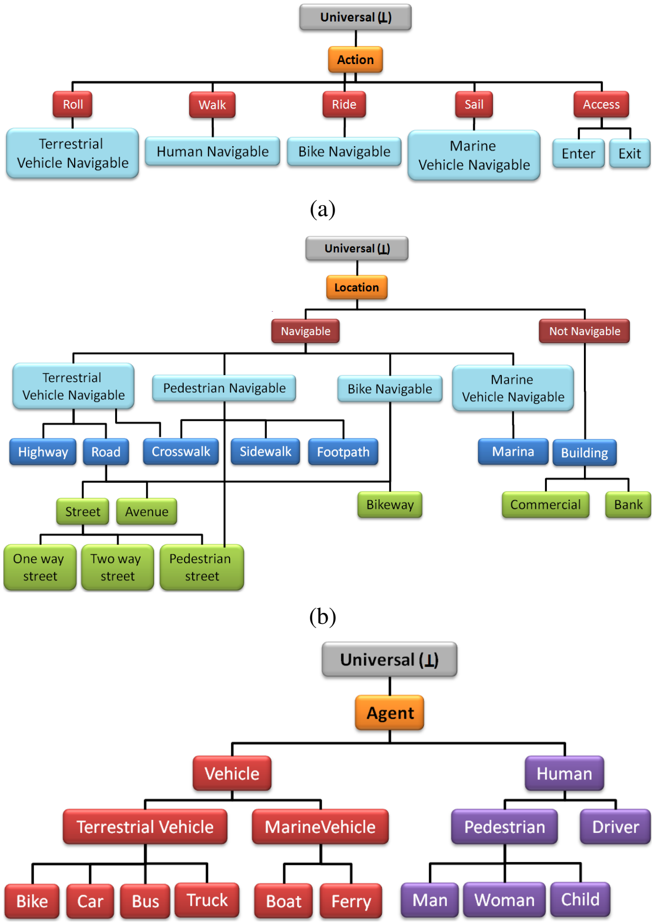
Taking advantage of the abstraction capabilities of the CGs formalism (through the *Concept Type Lattice* (CTL)) instead of representing different situated interactions of various agents in distinct locations, we are able to represent *abstract actions* performed by *agent archetypes* in *abstract locations*. Moreover, we first need to specify and characterise each of the abstract concepts. The concept type lattice enables us to specialise each abstract concept in order to represent situated behaviours such as path planning of agents in space. Figure 6 presents the first level of the concept type lattice refining the *agent*, *action*, and *location* concepts. Figures 7(a),7(b), and 7(c) present the expansion of the concept type lattice presented in Figure 6. Figure 7(a) illustrates some situated actions that can be performed by agents in the IVGE such as *sailing* for maritime vehicles, *rolling* for terrestrial vehicles, *walking* for humans, and *accessing* for humans to enter or exit buildings (we assume that buildings are not navigable locations from the perspective of outdoor navigation). 7(b) depicts how the *location* concept may be specialized into *Navigable* and *Not Navigable* concepts. The *Navigable* concept may also be specialised into *Terrestrial Vehicle Navigable*, *Pedestrian Navigable*, *Marine Vehicle Navigable*, and *Bike Navigable* which are dedicated navigable areas with respect to agent archetypes and environmental characteristics as specified by the *elementary semantics*. Figure 7(c) illustrates a few agent archetypes that are relevant to our geo-simulation including *pedestrians*, *cars*, *trucks*, and *bikes*.

In order to show how powerful such a representation may be, let us consider the following example. We want to build a MAGS simulating the navigation of three human agents (a man, a woman, and a child), two bike riders (a man and a woman), and three vehicles (a car, a bus, and a boat) in a coastal city. The navigation behaviours of these different agent archetypes must respect the following constraints (or rules): 1) *pedestrian* agents can only move on *sidewalks*, on *pedestrian street*s, and eventually on *crosswalks* if needed; 2) *vehicles* can move on *roads* and *highways*; 3) *boats sail* on the *river* and stop at the *harbour port*; and 4) *bikes* move on *bikeways*, *roads*, and *streets* but not on *pedestrian streets*. Using standard programming languages, it might be difficult to represent or develop the functions related to such simple navigation rules which take into account both the agents' and the locations' characteristics. However, the representation of these navigation rules becomes an easy task when using CGs and our defined concept type lattice. Here are their expressions in CGs:

[PEDESTRIAN:*p]<-(agnt)<-[WALK:*w1]->(loc)->[PEDESTRIAN NAVIGABLE:*pn]

[VEHICLE:*v]<-(agnt)<-[ROLL:*r1]->(loc)->[TERRESTRIAL NAVIGABLE:tn]

The arrows indicate the expected direction for reading the graph. For instance, the first example may be read: *an agent *p which is a "pedestrian" walks on a location *pn which is "pedestrian navigable"*. Since this expression involves the concepts *Pedestrian*, *Walk* and *Pedestrian Navigable*, this rule remains valid for every sub-type of these concepts. Therefore, thanks to CGs and the concept type lattice, there is no need to specify the navigation rules for men, women, and children if they act as pedestrians in locations such as *pedestrian streets*, *sidewalks*, or *crosswalk*. Indeed, these agent archetypes are subtypes of the *Pedestrian* concept and *pedestrian streets*, *sidewalks*, and *crosswalks* are subtypes of the *Pedestrian Navigable* concept. To conclude, CGs offer a powerful formalism to easily describe different concepts involved in MAGS including agents, actions, and environments.



(a)

(b)

(c)

Figure 7: An example of a conceptual description of *agents* archetypes (a), *actions* performed (b), and *locations* situated in a geographic environment (c).

V. Discussion

Thomas and Donikian proposed an Informed Hierarchical Topologic (IHT) [15] graph representing a part of the city of Renne (France) for human behavior animation purposes.

This graph is composed of three layers: (1) the *Basic Topological* layer which contains real urban objects modelled as simple spaces such as buildings and road sections; (2) the *Composite Space* layer which is composed of simple spaces or composite spaces of lesser importance; (3) the *Local Area* layer which is the highest level of the IHT-graph and which is composed of composite spaces. This hierarchical urban model allows manual abstraction of buildings into blocks and road-sections and crossings into roads. The abstraction process is done by the user which constrains and considerably limits its application to real world large-scale and complex geographic environments. Thomas's approach relies on a pre-defined decomposition of the virtual environment which is dedicated to urban environments. This decomposition is application-dependent (urban environments) and does not take into account the topologic and the geometric characteristics of the environment.

In contrast with Thomas [15] and Lamarche [8] approaches, our abstraction technique optimizes the representation of the geographic environment while taking into account the geometric, topologic and semantic characteristics of the geographic environment. This abstraction approach relies on an exact space decomposition technique (Constrained Delaunay Triangulation) in order to preserve the geometric and topologic characteristics of the geographic environment rather than on a pre-difined space decomposition. It also integrates semantic information associated with GIS data in order to enrich the description of the IVGE.

Embedding the information directly in the environment allows the support of agents' spatial reasoning capabilities. However, the preparation of the fully augmented geometric model is very time consuming and difficult due to the sheer amount of data. For example, a typical model of a city quarter as used by Farenc can contain several thousands of primitives of many types (such as polygons modeling sidewalk pieces, benches, trees, bus stops, etc.). Moreover, Farenc built the urban environment using data provided by Computer Assisted Graphic Design systems since the purpose of the simulation is computer animation. However, when building virtual geographic environments representing large-scale and complex geographic environments based on reliable GIS data, Farenc's approach can not be used since it is dedicated to exclusively represent urban environments. Indeed, the manual hierarchical space partitioning as proposed by Farenc is not feasible when dealing with geometrically complex environments. Moreover, the data structure of the urban environment's description as proposed by Farenc needs to be enhanced in order to manage a large amount of geometric and topologic data. Finally, the hierarchical structure should be built using the geographic environments's characteristics rather than being defined *a priori* as Farenc proposed.

The work done towards representation of semantic information in virtual environments has been mostly carried out at a geometric level [4]. Gutiuerez proposed a semantic model which aims to represent the meaning, and functionality of objects in a virtual scene [5]. However, since the purpose of Gutiuerez's approach is computer animations, the semantic information integration is located at the object description level rather than enriching the description the geographic environments. Virtual environments are usually created as computer graphics applications, with minimal consideration given to the semantic information [5]. Moreover, semantic information has been used in an *ad hoc* way without any standard formalism. There is a gap between geometry and semantic information in current virtual geographic environment models. Since we believe that semantic information integration into a VGE's description is by nature a knowledge representation problem, a suitable and standard knowledge representation formalism has been proposed to integrate semantic information in the VGE's description.

VI. Conclusion and Future Works

In this paper, we introduced our IVGE model which automatically buids semantically-enriched and geometrically-accurate description of informed virtual geographic environments. We also proposed an abstraction approach of the IVGE's description in order to support large-scale and complex geographic environments. First, we described a ***geometric abstraction*** process which enriches the IVGE description with terrain semantics. Moreover, the geometric abstraction process helps to detect and filter elevation anomalies and qualifies the terrain shape, specifically slope. Second, we detailed a ***topologic abstraction*** which builds an hierarchical topologic graph in order to deal with large-scale virtual geographic environments. This hierarchical structure reduces the size of the topological graph representing the IVGE. Third, we showed how the ***semantic abstraction*** process enhances the hierarchical topological graph using the concept type lattice in order to build different views of the IVGE. We are currently working on the leverage of our enhanced IVGE model to support hierarchical path planning algorithms which take into account both the abstracted description of the IVGE and the agent type's characteristics.

Acknowledgement

M. Mekni benefited from a PDF scholarship granted by FQRNT (***F****onds* ***Q****uébécois de la* ***R****echerche sur la* ***N****ature et les* ***T****echnologies*). The author thank Dr Froduald Kabanza, Univ. Sherbrooke, Qc for his advice during this research.

References

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

[2] S.-G. Chen and J.-Y. Wu. A geometric interpretation of weighted normal vectors and its improvements. In *Proceedings of the International Conference on Computer Graphics, Imaging and Vision: New Trends*, pages 422–425, Beijing, China, 26-29 July 2005.

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

[4] A. Garcia Rojas Martinez. *Semantics for Virtual Humans*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2009.

[5] M. Gutérrez Alonso. *Semantic Virtual Environments*. PhD thesis, École Polytechnique Fédérale de Lausanne,Swisse, 2005.

[6] M. Kallmann. *Object Interaction in Real-Time Virtual Environments*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2001.

[7] M. F. Klügl and M. Neumann. Landscape abstractions for agent-based biodiversity simulation. In *Fourth International Joint Conference on Autonomous Agents & Multiagent Systems*, Utrecht University, The Netherlands, July 2005.

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

[9] J. Lewis, P. Skarek, and L. Varga. A rule-based consultant for accelerator beam scheduling used in the CERNPS complex. In *ICALEPCS'95: International Conference on Accelerator and Large Experimental Physics Control Systems*, Chicago, Illinois USA, October 30-November 3 1995.

[10] M. Mekni and B. Moulin. Holonic modelling of large scale geographic environments. In *HOLOMAS'09: 4th International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, Linz, Austria, September 2007.

[11] M. Minsky. *A Framework for Representing Knowledge*. MIT-AI Laboratory, 1974.

[12] S. Paris, S. Donikian, and N. Bonvalet. Environmental abstraction and path planning techniques for realistic crowd simulation. *Computer Animation and Virtual Worlds*, 17:325–335, 2006.

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

[14] J. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Course Technology, August 1999.

[15] R. Thomas and S. Donikian. A model of hierarchical cognitive map and human memory designed for reactive and planned navigation. In *Proceedings of the 4th International Space Syntax Symposium*, volume 1, pages 72–100, Londres, 2003.

[16] S. E. Varlan. Knowledge representation in the context of e-business applications. *BRAND. Broad Research in Accounting, Negotiation, and Distribution*, 1(1):1–4, September 2010.

Systems of Systems Concept in Knowledge Management

Alexander Petrovich Kamyshanov
Accounting and Audit Department
Plekhanov Russian Economic University
Moscow, Russian Federation
e-mail: a_kamyshanov@rbcmail.ru

Peter Ivanovich Kamyshanov
Accounting and Audit Department
Plekhanov Russian Economic University
Moscow, Russian Federation
e-mail: a_kamyshanov@rbcmail.ru

Abstract — **The development of new acquisition technologies is important function of knowledge management. They are realizing via powerful computer tools such as the Internet interactive hypermedia or the Large Knowledge Collider. But except powerful tools new cognitive concepts and procedures should be synthesized for knowledge evolution. The Systems of Systems approach gives the possibility to design knowledge bases network. In the framework of this concept meta-level knowledge base construction procedures are described for some problems solutions.**

Keywords-Systems of Systems; knowledge acquisition technologies

I. INTRODUCTION

Knowledge management (KM) appeared and developed as the new research area in computer science in dialectical conjunction with philosophy, epistemology, and theory of management. In modern economy KM is based on information technologies and cognitive science. Key concepts of KM are knowledge and education that can be treated as the business educational and innovative intellectual products and services, which can be transferred for a high value return. Otherwise they are looked as the productive assets for high competitive enterprise. Research methods of KM such as the construction of ontology for representing the main categories and entities in particular scientific domain, the rules formalization for logic inference to acquire new knowledge, and others give the possibility to get the new insight on already formed fundamental knowledge bases of the physics and the chemistry. The design of models for the knowledge evolution and development is one of the items in knowledge management. Spreading Piaget cognitive development theory for knowledge bases the two components model was proposed in [1]. One component is presented by ever changing content and structures as the other is realized by unchangeable functions.

II. THEORETICAL BACKROUND AND TOOLS FOR KNOWLEDGE ACQUISITION

The discovering of sources for stable and unchangeable true knowledge with the aim to incorporate them into ontology is one of the main problems in knowledge management. Inherently concepts in every area of research resulted democratic discussions of scientists with equal rights for the truth. However, every time it is not possible to consider the positions of different scientific schools via creation stable and unchangeable functions in ontology. Next example illustrates this thesis. In financial management and accounting the main categories and entities are defined by national legislations and standards. On the contrary in nature sciences such as the physics and the chemistry there are now laws and rules signed by the head of one or another state. Analysis has shown that in these cases most truthful, reliable, and confirmed by experiments knowledge are published in the textbooks for universities and encyclopedias. This statement is based as on multistage reviewing processes, numerous qualified readers, and free access for upper pointed publish sources, so on the great editors' responsibility before the future generations of students and researchers. No one can deceive the future.

On other side, in many problem domains ever changing content and structures of the ontology are representing by e-textbooks and various Wiki systems in the Internet hypermedia space [2]. The characteristic feature of these knowledge acquisition tools is the cooperative creation of new knowledge and paradigms by realizing every user possibility to change the content and items relations.

In the Internet hypermedia space knowledge are coming through Socialization, Externalization, Combination, Internalization (so-called SECI-process), which consider a spiraling evolution interaction between explicit and tacit knowledge resulted by their synergy [3]. But the young students with non formed stable knowledge component can be overloaded with information.

Following the new knowledge development paradigm, in which everyone is creating hypermedia content and information for Wikipedia and other bases, a learner or researcher has millions of information pieces at his fingertips varying in quality and relevance to the actual scientific work. Great amount of information may lead to the inability to discern between facts and concepts. The reason is the significant number of conflicting positions. Nevertheless the appearance of new Internet tribune for scientific discussions was the greatest achievement in democratization of science that accelerates the knowledge evolution. New research projects and programs are developing with the aim to raise the efficiency of knowledge acquisition systems. In European Union (EU) Framework Program 7 manages and coordinates the activity for creation the Large Knowledge Collider (LarKC). It is determined as

ISBN: 978-1-61208-121-2

a platform for massive distributed incomplete reasoning directed to remove the scalability barriers of currently existing systems for Semantic Web [4].

So in our days two approaches for knowledge acquisition and development are realizing simultaneously. One is based on the Internet interactive multimedia hyperspace, while other is presented by LarKC technologies.

Thus the procedures synthesis for acquisition new knowledge by utilizing the potentials of LarKC and the Internet becomes extremely important item of agenda. Epistemology methods and the history of science have demonstrated that new knowledge appears as the results of hard experimental work that causes the evolution of insight within the problem domain of particular science. Otherwise the emergence of new paradigm gives the possibility to develop new theories. In some sciences, for example in mathematics, new knowledge results the hard theoretical work with already existing knowledge bases.

III. Systems of Systems Concept for Knowledge Bases

A. *Systems of Systems Engineering*

Every science can be looked as a complex system composed of subsystems. Each is characterized by hierarchy of interacting and networking components formed by concepts, laws, entities, and atomic terms described in ontology and united via logic inference constrains. Knowledge system of particular science is operating and coordinating as in conjunction with multiple objectives, so to evolutional perspectives. In the physics a hierarchy or continuum of laws as distinct systems or disciplines that are cooperating and interdependent was investigated in [5]. In biology system approach was introduced at the end of 1960th decade by the fundamental works [6][7]. The philosophy of systems engineering applied to real world had been developed in [8].

Traditionally every science was the logically separate knowledge system with little interdependence with others. The progress in knowledge management and information technologies is rapidly changing our world. Close relationship of different sciences and their interactions are the topics of current agenda.

Systems of Systems (SoS) engineering (SoSE) and analysis had appeared and developed in conjunction with space operations as well as large scale information systems and software design for them [9][10]. Research envisioned for SoS includes investigation oriented to the creation of new systems engineering methodologies to cope with SoS evolution and emergent knowledge associated with it.

Meta-model design is one of the powerful tools worked out in the framework of systems research [11]. Creatively applying it to SoSE for knowledge management we can describe the problem in the form of monotonous knowledge base for particular scientific domain.

Knowledge base KB constructed from logic declarations is considered monotonous, if for every declaration α and β the following statements are correct:

$$\text{if } KB \mid= \alpha \ \text{ then } KB \wedge \beta \mid= \alpha . \qquad (1)$$

In the case when declaration β breaks the KB monotonous property, in the ontology this declaration is considered as the new atomic term. The no contradiction logic is the main property for every scientific problem domain. It is ensured by collective efforts of researchers working all over the world in the various sciences.

As the problem knowledge base has been constructed, new scientific domains necessary for the solution should be investigated. In these domains monotonous segments of knowledge bases related to the problem should be determined. After that all segments are united in new monotonous meta-level knowledge base. Finally solution of the problem can be defined in the form of theorem proving by resolution method.

To illustrate the knowledge evolution model for the chemistry and the physics we can look at the category of the atomicity or the valence. The atomicity is important as for creation of semiconductor devices, so for determination the molecular structure of chemicals. It appeared at 1425 in the framework of medieval hermetic art. Later this category got the development in the chemistry and in the physics. In our days valence has found final formalization in quantum mechanics, where complex mathematical instruments and semantic phenomenological Pauli principle define it [12]. In this case, Pauli principle may be considered as an example of construction the meta-level knowledge base for description the atomicity properties.

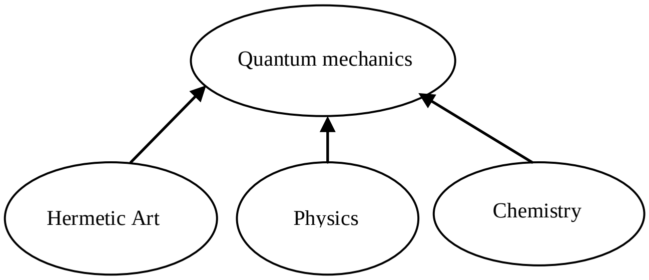


Figure 1. Knowledge bases network for definition the atomicity.

At the same time most other alchemy concepts modern chemists and physicists regard as pseudoscience.

B. *Systems of Systems Knowledge Engineering in Cardio Physiology*

There is one more example. At 1560th in Spain the great Renaissance doctor Andreas Vesalius tragically had discovered that human heart could operate independently from other systems of organism. Nevertheless, in this case the laws of hydromechanics for blood flow had to be realized.

In quite mode the minute blood flow volume over the man healthy heart is about $3 \sim 4 \times 10^{-3}$ m^3. The blood density *(ρ)* is approximately 1050 ~ 1064 kg/m^3. The frequency of the cardiac rate at quite mode is near 72 strikes per minute. Under hard physical load the frequency of the cardiac rate may increase till 210 strikes per minute and the minute blood flow volume rises up to 40×10^{-3} m^3. The square of the aortic valve, over which blood leaves the heart, is about 3×10^{-4} m^2. If the circle with the same square approximates valve's form, then the diameter of the circle d will be equal

$$d = 2 \times (S / \pi)^{0.5} \approx 2\times10^{-2}\,m\,. \qquad (2)$$

Under this condition the blood flow (*BF*) over heart equals 3.6×10^{-3} m^3 and the initial velocity of the flow V_0 will be

$$V_0 = BF / S = 3.6\times10^{-3} / 3\times10^{-4} = 1.2\ m\ p.s. = 720\ m\ p.\ h. \ (3)$$

For the minute blood flow volume equals 40×10^{-3} m^3, V_0 will be 8000 m p. h.

In the vessels blood liquid flow is considered to be continuous. At the aorta, by which arterial vascular system begins, the mean value of the hydrodynamic pressure for healthy heart is about 100 mm hg. col. According cardiology data in the right heart atrium, where venous loops are ending, the hydrodynamic pressure approximately equals 0 mm hg. col.

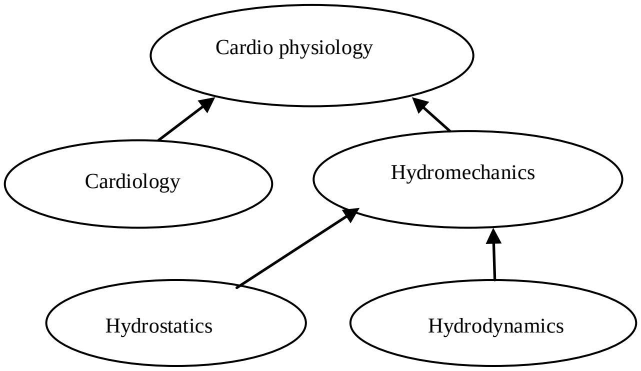


Figure 2. Knowledge bases topology for venous loop hydromechanics in blood circulatory system.

In [13] the theorem had been proved that in the venous loop of blood circulatory system the flow's pressure is defined by the laws of hydrostatics realizing in pulse mode.

As the consequence from the theorem, the continuity of blood flow over different organs in the body is provided by the venous valves operating in the definite way. Errors in the algorithm controlling the valves activity or their damage can cause the diseases of venous system, the heart failure, and trophic ulcer.

IV. Meta Level Knowledge Base as Hidden Variable

A. Knowledge Base Engineering for Brain-Computer Interface

At the beginning of 1990th IBM, Hewlett Packard, and Sony corporations started the sales of color monitors for the personal computers. These devices experimentally and publicly had demonstrated that human nerve fiber in optic track can transmit visual information (VI) with bit rate more than 135 109 × 10^{12} bit/sec.

There are 1280 pixels over monitor horizontal and 1024 over vertical. Machine word in 32 bits executes the colors and brightness code. It is known from physiology, that more than 24 Hertz frame frequency is required to create illusion of moving picture. Thus

$$VI \geq 24 \times 1280 \times 1024 \times (2^{31} + 2^{30} + 2^{29} + ... + 2^{0}) \approx$$
$$\approx 31\,457\,280 \times 4.3 \times 10^{9} \approx 135\,109 \times 10^{12}\ bit/sec. \qquad (4)$$

From physics is known, that visual signals are transmitting in frequency spectrum spreading from 4×10^{14} till 8×10^{14} Hertz [14][15]. In common case healthy man and woman can get visual information traffic with lowest probability magnitudes of the first and second type errors.

At the same time, in [16] was pointed out that current brain–computer interface (BCI) had maximum information transfer rate up to 10 - 25 bits/min. This limited capacity could be valuable for people, whose severe disabilities prevent them from conventional communication method. There were declared, that future progress in multimedia neurophysiology interface development will be possible via involving neurobiology, engineering, mathematics, and computer science.

It is interesting fact that significant number of neurophysiologists has quite reliable visual systems providing lowest probability magnitudes of the first and second type errors for visual perception. Nevertheless they are persisting in following the concepts that nerve signals are transmitting by electrical impulses with voltage in few mill volts and maximum frequency in 500 Hertz. There is no one computer or telecommunication engineer, who can provide reliable transmission of the electrical signals with such parameters.

To solve this problem in [17] the next network with meta-level knowledge base as hidden variable was proposed [18].

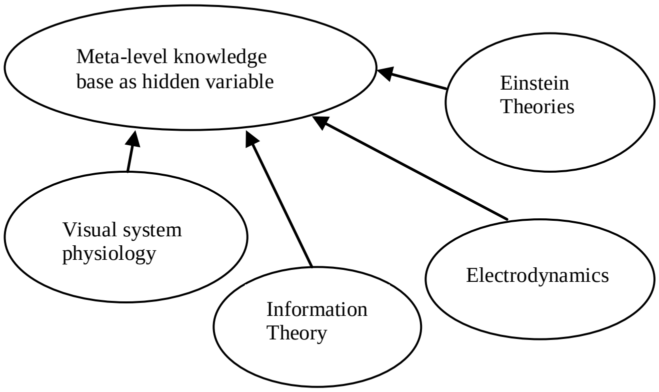


Figure 3. Cognitive model for visual system with meta-level knowledge base as hidden variable.

The investigation of the Internet hypermedia space had shown that to overcome the emergent difficulties new knowledge bases of biochronotopology and psychronotopology should be executed as hidden variable [19]. Concepts and categories defining by these knowledge bases could be very important for new research.

Examination of bio chronotopology category had led to new problems.

B. The Problems of the "Time" Category

Three dimensions of space and one dimension time had formed human perception of environmental world. "Time-dimension" is directed from past to present and from present to future. It is regarded, that such "time" direction is determined by "cause-effect" relation between events in physical world [20][21]. At another side the physical approach to the "time" category is not predetermined. In this science the "time" category has several different meanings.

First of all, "time" is looked at as independent continuous parameter in mathematical equations describing evolutions of physical systems.

The next, "time" is presenting the scale for determination the consequence of physical events. There are several types of such scales utilizing now. Astronomic scale is the main one. It defines the time intervals through the consequence of astronomic events, such as the rise and the setting of the Sun, consequent, following one after another two culminations of the observed star over the Earth particular meridian. Produced on the base of quantum frequency generators atomic "time scales" are also wide spread.

Additional meaning of the category "time" was realized in quantum mechanics, where the "fifth", the "sixth" and so on "time-dimensions" were introduced for "space-time continuum" [22][23]. These additional "time parameters" were orthogonal to traditional linear "time", and so the attempts were done to overcome the quantum uncertainty relation. However, as this aim had been reached, new approach had born out uncertainty in Einstein relativity theories [24]. Under the condition of "three-dimensional time" there wasn't presented any formula for the speed of light c determination. The speed of light c is considered as one of the fundamental physical constants. For authors' opinion, these scientific ideas have a great potential for development [25].

While in mathematical equations "time" is continuous variable, "time" determined by various physical scales has a discreet property.

Quantum physics characterizes atom as the system composed by nucleus with electrons' shells. Atomic systems can be united into complex molecular systems (Systems of Systems) existing in volatile SoS environment.

If the atomic quantum frequency standard determines the category "time" in molecular SoS, then serious problems can appear in description of molecular SoS evolution. Following N. Bohr postulates the frequency of the emitting radiation υ is equal to the quantity $(E_i - E_j) / h$ [26].

Here E_i , E_j - are the total energies of the stable electron's orbits, h – is Planck constant.

Under such approach, for physical "time" scale determined by atomic quantum frequency generator is impossible to define the "time" parameter characterizing evolution of the electron in atom at the stable energy state. Following modern quantum mechanics paradigm in the coordinate system connected with this electron it has neither past, nor future only present continuous. Moreover, in atom, for electron existing on the stable energy level the "cause-effect" relation is undetermined. In our days there are no technical tools defining category "physical time" for electron in a stable atom system.

Physical indetermination of "time" parameter for stable electrons' orbits put forward the question about physical meaning of parameter t in Schrödinger equation for wave function $\Psi\ (x, y, z, t)$ [27]

$$(- \frac{\hbar^2}{2m} \times\Delta + W(x, y, z, t)) \times \Psi(x, y, z, t) = i \times \hbar \frac{\partial}{\partial t}\ \Psi\ (x, y, z, t). \quad (5)$$

Here Δ - is Laplace operator, x, y, z, m – are electron's coordinates and mass, $W\ (x, y, z, t)$ – is the potential energy of electron in atom, $\hbar = h\ /(2\times\ \pi);\ i = \sqrt{-1}$.

Chemical elements with significant atom numbers and protein macromolecules with complex three-dimension architecture have electrons' shells with complex topology. Schrödinger equations should be integrated for them [28]. The problem is that in modern textbooks for universities the "time" scales construction procedures for quantum objects don't determine. In control science this case is describing by fuzzy system paradigm [29][30][31].

V. Perspectives for Knowledge Bases Systems of Systems Concept

A. Knowledge Management for Bio Cybernetics

Bio cybernetics had demonstrated the possibility to model the complex biological system activity on the base of physical, chemical, hydromechanics' laws and the theory of automatic control. The results exhibited the experimental proofs for modeling human blood circulation at the definite range of environmental parameters were published in [32] [33][34] at 1970. There were shown that circulatory system dynamic could be modeled by the system of linear differential equations

$$\dot{X} = A\ (t) \times X + B \times u\ . \quad (6)$$

The research work for modeling cardio system activity continued in the next decades [35][36]. The electro-chemical model for the cardiac cell utilizing the system of differential equations with more than 50 parameters was proposed in [37]. The software tool to simulate cardiac cell activity was described and demonstrated in [38].

But as 40 years had passed no experimental results were published concerning the biological organs or the systems of any animal or human providing the integration of differential equations demanded for the blood circulation control.

The problem is extremely important for modern physiology. Several hundred millions of people are suffering from diseases of heart and circulatory system. To overcome the emergent difficulties the knowledge bases network is proposed, see Figure 4.

For authors' opinion, in nature the control algorithm for blood circulatory system is constructed on the base of fuzzy rules [30] in conjunction with multilevel qualitative

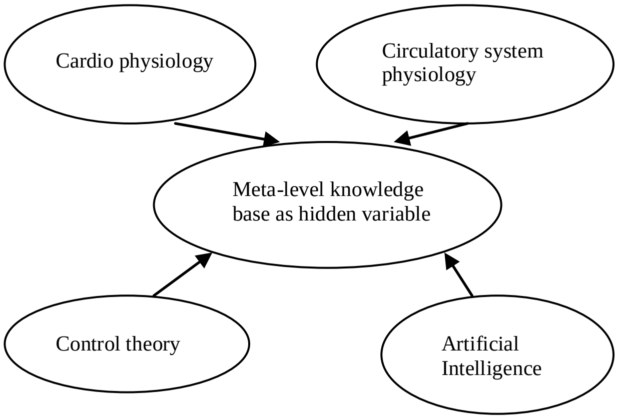

Cardio physiology
Circulatory system physiology
Meta-level knowledge base as hidden variable
Control theory
Artificial Intelligence

Figure 4. Cognitive model for circulatory system control with meta-level knowledge base as hidden variable.

reasoning trees [39] and qualitative induction logic inference [40]. Control signals are realizing as by bioelectrical impulses so by biochemical agents. With sufficient details they are described in the textbooks for medical universities.

B. The Elementary Particles Physics Problems

One more fundamental problem is studying in the modern textbooks for the chemistry and the physics. It is known from these sources that as hydrogen atom so neutron can be split on proton and electron. But while a hydrogen atom is considered to be a stable physical or chemical system, on the contrary, a free neutron has half life period about 10 minutes. The problem is that the properties of hydrogen atom are investigated by the chemistry or by the atomic physics, while properties of neutron are the subject of the elementary particles physics. This case also can be described by the network with monotonous meta-level knowledge base as hidden variable.

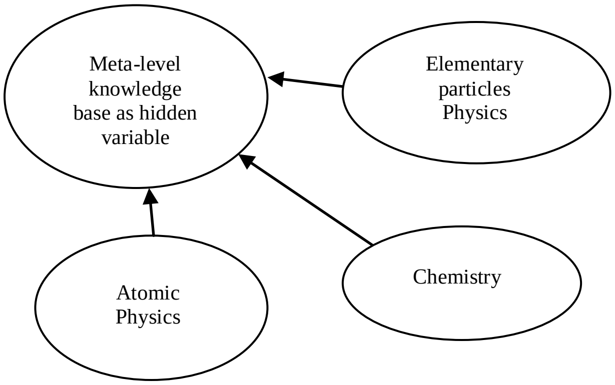

Meta-level knowledge base as hidden variable
Elementary particles Physics
Chemistry
Atomic Physics

Figure 5. Cognitive model for hydrogen atom stability.

It was written in [41] that over the 20th century electronics had exhausted completely the atom model consisting of electrons orbiting the nucleus. New ideas are wanted to improve the matter structure model. Modern powerful computer tools such as the Internet interactive hypermedia space and LarKC can be utilized in the framework of SoSE concept to solve this problem. Scientific community has the real opportunity to estimate experimentally the efficiency of these acquisition tools and approaches.

New computer technologies provide the possibility to get new insights, which are determining as the new type of knowledge. So they give a chance to modify the current theoretical bases in various sciences. New insights have to be represented in explicit symbolic form [42]. Therefore, for authors' opinion, the Internet interactive hypermedia space is more suitable for discovering new scientific concepts than LarKC operating in the frame of monotonous logic formalism.

VI. Conclusion and Future Work

In general knowledge acquisition should be resulted by the development of new theories and their public recognition. Thus SECI processes will get their logical final. Modern computer technologies give the possibility to accelerate them. Systems of Systems engineering for construction the new knowledge bases can be applied for complex problems yet not solved.

In knowledge management SoSE concept has not only the advantages but the difficulties as well. For example, there are very few cardiologists realizing that the existence of the correct elements is not sufficient condition for reliable, correct operation of the control system. Stable functioning automatic control system can be designed from unstable units. It is necessary to know that blood circulatory system is referred to nonlinear, multi loops control systems, in which auto oscillation and sliding modes can exist. Under these conditions the stable operation of the cardiovascular system is determined not by electrocardiogram signals registration, the arterial pressure and pulse frequency measurement but by the methods of control theory.

On the other side, no more than one hundred researchers in the automatic control domain possess the knowledge about the structure of blood circulatory system, the peculiarities of the arterial and the venous vascular loops, about biophysical and biochemical processes in the blood. So in the case, when, for example, high level ontology in cardiology with great efforts and expenses will be created, none scientist will be able to work with full-scale version. To great concern, this statement is true for the problems of the Elementary particles physics and the fuzzy logic systems.

Practice has shown, the collective research work is the most effective way for unification the ontology of different sciences. As the result, new knowledge management methods will appear. These methods must be published in the textbooks for universities.

Now Systems of Systems engineering is intensively executed in financial management. For authors' opinion, in this area international cooperation will be very fruitful.

References

[1] C. Stickel, M. Ebner, and A. Holzinger, "Useful oblivion versus information overload in e-learning elxamples in the context of Wiki systems," Journal of Computing and Informational Technology, vol. 16, Dec. 2008, pp. 271-277.

[2] H. Maurer, “Web-based knowledge management,” IEEE Computer, vol. 31, (3), 1998, pp. 122-123.

[3] I. Nonaka and H. Takeuchi, The Knowledge Creating Company. Oxford: Oxford University Press; 1995.

[4] M. Witbrock, “Acquiring and using large scale knowledge,” Proc. 32nd Int. Conf. on Information Technology Interfaces (ITI2010), Univ. Comp. Centre, Univ. of Zagreb, Jun. 2010, pp. 37-42.

[5] R. Feynman, The Character of Physical Law, Cambridge, Massachusetts: MIT Press, 1965.

[6] M. Mesarovic (Ed.), Systems Theory and Biology, New York: Springer-Verlag, 1968.

[7] M. Mesarovic, D. Macko, and Y. Takahara, Theory of Hierarchical, Multilevel Systems, New York: Academic Press, 1970.

[8] A. Gheorghe, Applied Systems Engineering, New York: John Wiley & Sons, 1982.

[9] A. Sage, Systems Engineering, New York: John Wiley & Sons, 1992.

[10] A. Sousa-Poza, S. Kovacic, and Ch. Keating, “System of Systems engineering an emerging multidiscipline,” Int. J. System of Systems Engineering, vol. 1, Jan. 2008, pp.01–17.

[11] A. Hall III, Metasystem Methodology: A New Synthesis and Unification, New York: Pergamon Press, 1989.

[12] J. Bell, Speakable and Unspeakable in Quantum Mechanics, Cambridge: Cambridge University Press, 1987.

[13] A. Kamyshanov, “Semantic model design for veins’ loop by the logic formalism of Artificial Intelligence,” Proc. 16th IASTED Int. Conf. on Applied Modelling and Simulation (AMS2007), ACTA Press, Aug. 2007, pp. 336-341.

[14] J. Seleznev, The Bases of the Elementary Physics, Moscow: Nauka, 1974.

[15] S. Palmer, Vision Science: Photons to Phenomenology, Cambridge, Massachusetts: MIT Press, 1999.

[16] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T.Vaughan, “Brain-computer interfaces for communication and control,” Clinical Neurophysiology: Official Journal of the International Federation Neurophysiology, vol. 113, Jun. 2002, pp. 767 – 791.

[17] A. Kamyshanov, “Meta-model design for Internet multimedia neurophysiology interface,” Proc. of 3rdAMS2009, Asia International Conference on Modelling and Symulation (AMS2009), IEEE Com. Societ. CPS, May 2009, pp. 632-636.

[18] S. Russell, J. Binder, D. Koller, and K. Kanazawa, “Local learning in probabilistic networks with hidden variables,” Proc. of 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal: Morgan Kaufmann, 1995 pp. 1146-1152.

[19] A. Dubrov, Symmetry of Biorythms and Reactivity, New York, London, Paris: Gordon and Breach Science Publishers, 1988.

[20] J. Dorling, "The Dimensionality of time," American Journal of Physics, vol. 38, (4), 1970, pp. 539-540.

[21] Yu.Vladimirov, N. Mitskievich, and J. Horsky, The Space, Time, and Gravitation, Moscow: Mir Publishers, 1987.

[22] F. Tangherlini, "Atoms in higher dimensions," Nuovo Cimento, vol. 14, (27), 1963, pp. 636-651.

[23] Xiadong Chen, “A New Interpretation of Quantum Theory | Time as Hidden Variable,” Department of Physics, University of Utah, Salt Lake City, UT-84112, 1999. Available on: *http://authors.aps.org/eprint/files/1999/Feb/aps1999feb10_004/auxiliary/New_Interp.ps*

[24] A. Einstein, The Collected Scientific Works, v. 1-4, Moscow: Nauka, 1965-1967.

[25] R. Penrose, The Large, the Small and Human Mind, Cambridge: Cambridge University Press, 1997.

[26] N. Bohr, The Three Articles about Spectrums and Atom Structure, translated from German, Moscow, 1923.

[27] Available on: http://hyperphysics.phy-astr.gsu.edu/HBASE/quantum/

[28] M. Lockwood, Mind, Brain, and the Quantum, Oxford: Basil Blackwell, 1989.

[29] L. Zadeh, Fuzzy Sets and Application, New York: John Wiley & Sons, 1987.

[30] R. Yager and L. Zadeh, An Introduction to Fuzzy Logic Applications in Intelligence Systems, New York: Kluwer Academy Publishers, 1991.

[31] L. Zadeh, “Fuzzy logic, neural networks, and soft computing,” Comm. ACM, vol. 37, (3), 1994, pp. 77-84.

[32] B. Petrovskij, V. Shumakov, V. Novoselcev, E. Shtengold, V.Baikovskij, L. and A. Dartau, “Vital important human organism functions providing is the base of Artificial Heart automatic control,” Proc. IFAC 1st Int. Symp. Technological and Biological Problems of Control, 1968, Erevan, USSR, Bioelectrical Control. Man and Automatic Systems, (Moscow: Nauka (Science), 1970, (in Russian), pp. 278-287.

[33] J. Krasner, P. Nardella, and P. Voykydis, “Cybernetic problems of the device for physiological control of Artificial Heart,” Proc. IFAC 1st Int. Symp. Technological and Biological Problems of Control, 1968, Erevan, USSR, Bioelectrical Control. Man and Automatic Systems, (Moscow: Nauka (Science), 1970, (in Russian), pp. 265-275.

[34] W. Pickering, P. Nikiforuk, and J. Merriman, “The Use of analog computer for analysis of control mechanism in cardio vascular system,” Proc. IFAC 1st Int. Symp. Technological and Biological Problems of Control, 1968, Erevan, USSR, Bioelectrical Control. Man and Automatic Systems, (Moscow: Nauka (Science), 1970, (in Russian), pp. 408-421.

[35] C. Luo and Y. Rudy, “A Dynamic model of the cardiac ventricular action potential,” Part I, II, Circulation Research, vol. 74, (6), 1994, pp. 1071-1113.

[36] Y Lecarpentier, C. Coirault, and D. Chemla, “Regulation cellulaire et moleculaire de la contruction cardiaque,” Medecine Therapeutique, vol. 2, (2), 1996, pp. 113-122

[37] R. Roche, R. Lamanna, M. Delgado, F. Rocaries, and Y. Hamman, “Calcium homestasis and membrane potential in cardiac myocite: an electrochemical model,” Proc. of 5th EUROSIM Congress on Modelling and Simulation, France, Paris, 2004.

[38] R. Roche, R. Lamanna, M. Delgado, F. Rocaries, Y. Hamman, and F. Peker, “A Software tool for the simulation of a cardiac cell,” Proc. 16th IASTED Int. Conf. on Applied Modelling and Simulation (AMS2007), ACTA Press, Aug. 2007, pp. 336-341.

[39] B. Kuipers, Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge, Cambridge, Massachusetts: MIT Press, 1994

[40] D. Šuc, “Machine reconstruction of human control strategies,” Frontiers Artificial Intelligence Appl., vol. 99, Amsterdam: IOS Press, 2003.

[41] D. Packard, The HP Way. How Bill Hewlett and I Built Our Company, Cambridge, Massachusetts: Harvard Business School Press, 1995.

[42] I. Bratko, “An Assesment of machine learning methods for robotic discovery,” Journal of Computing and Informational Technology, vol. 16, Dec. 2008, pp. 247-254.

Modeling of Microsystems Production Processes for the *MinaBASE* Process Knowledge Database Using Semantic Technologies

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

***Abstract*—In this paper, we present the consolidation of a process knowledge database for knowledge-intensive production processes in the field of microsystems technology with a workflow component. Among the requirements to be met by the workflow component are the hierarchical presentation of process chains, a close integration of the product structure in the form of assemblies, modules, and components, storage of previous (unsuccessful) attempts together with the information arising, the derivation of process patterns from concrete workflow instances, and the explicit modeling of dependencies among various steps in the workflow. In contrast to existing workflows, the complexity in the concrete microsystems technology application does not lie in potential branchings of activities, but in the inherent information and concepts of the process knowledge database and their relations and constraints. Starting from the metamodel developed for process modeling by List and Korherr, a multi-perspective model with four overlapping and integrated perspectives (system, process, project, and development perspectives) was developed to better manage the complexity of and reuse individual knowledge entities. As a proof of concept, the model is implemented by means of formal knowledge representation languages from the semantic web, which will be illustrated using the previously analyzed development perspective as an example.**

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

I. Introduction

Knowledge, experience, and capacities of the employees make up the core competencies of an enterprise and have a crucial influence on its competitiveness. The knowledge required for creating values added is no public good, but a business resource that has to be administrated efficiently in order to ensure economic success. For software-technical support, knowledge management systems [1] have been established. In process-oriented knowledge management [2] highly knowledge-intensive production processes in microsystems technology, for example, are managed. Production processes in this field are characterized by a high interdisciplinarity, many process steps, and a low standardization. Frequently, a product is produced by an individually tailored production process [3]. Moreover, design decisions during the development of microsystems are strongly dependent on the characteristics of the different fabrication technologies applied. They may require knowledge of various disciplines like microoptics, biotechnology, or sensor technology. In practice, technical experts have to deal with heterogeneous, distributed, and partly incomplete data, such that new or already solved product development problems are difficult to handle. Knowledge management methods represent an promising approach to overcoming this barrier. The Institute for Applied Computer Science has developed the *MinaBASE* process knowledge database which structurally acquires the expert knowledge of microsystems technology and makes this knowledge available centrally, homogeneously, and collaboratively [4]. However, the *MinaBASE* approach does not provide for any process-oriented linking of these structured knowledge entities. Linking would allow not only for the representation of the knowledge required for microsystems production processes, but also for the modeling of these processes on an abstract level. So, an approach to process-oriented linking of existing process data will be presented in the following sections.

The paper will be structured as follows: The next section will present the underlying process knowledge database *MinaBASE*. Then, requirements made on process modeling in this context will be highlighted and existing process modeling standards will be analyzed for suitability. A solution approach will be described and implemented using semantic technologies to derive implicit knowledge from the modeled facts. A part of modeling as ontology will be described in detail in Section VI.

II. *MinaBASE* Process Knowledge Database

The *MinaBASE* process knowledge database was developed within the framework of the MikroWebFab joint project funded by the BMBF [2]. Technology partners

ISBN: 978-1-61208-121-2

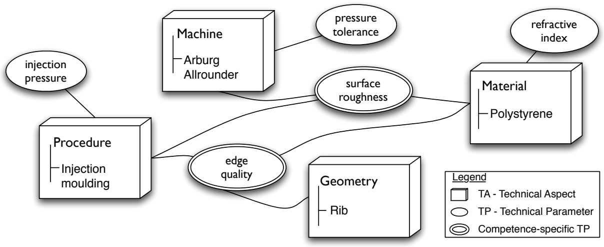


Figure 1. Schematic representation of a *MinaBASE* competence

of a virtual enterprise used it for the structured storage of technical production parameters of the processes and materials used in microsystems technology and of partner-specific technical competencies. The so-called technical aspects (TA) that serve to model materials, machines, and fabrication technologies are the smallest information entity in MinaBASE [5]. TA are arranged in taxonomies using generalization hierarchies. The number and contents of taxonomy trees can be specified and modified during runtime, such that a flexible structure tailored to meeting the requirements of microsystems technology can be defined for the storage of production knowledge. TA can be assigned properties that are referred to as technical parameters (TP). A TP is specified as a character string, integer, or floating-point number and references an attribute, e.g., density. As in the object-oriented approach, the TP of a TA are passed on to partial hierarchies located below in the hierarchy tree. In addition, lower hierarchy levels can further refine the inherited TP by specifying general value ranges.

To model the capacities of a technology partner, competencies [4] are considered a set of various TA of disjunct hierarchy trees, which is illustrated in Figure 1.

Here, the competence of ”injection molding of a polystyrene web using the Arburg Allrounder machine”, with several TP is represented schematically. The respective TA are selected from the hierarchy trees of process, machine, material, and geometry element, with the TA having own TP, such as ”injection pressure” of the injection molding process. Combination of these TA yield the competence having other TP, such as the edge quality and surface roughness. Consequently, a competence is a type of view on a certain combination of TA with properties in the form of TP that are only valid for this combination and, hence, characterize the competence in more detail. Accordingly, TA can be used in several competences, which illustrates their role as reusable, encapsulated, smallest information entity.

III. Requirements

In this section, the boundary conditions of and requirements on process modeling for *MinaBASE* shall be analyzed. Five types of requirements can be distinguished and will be described in more detail below.

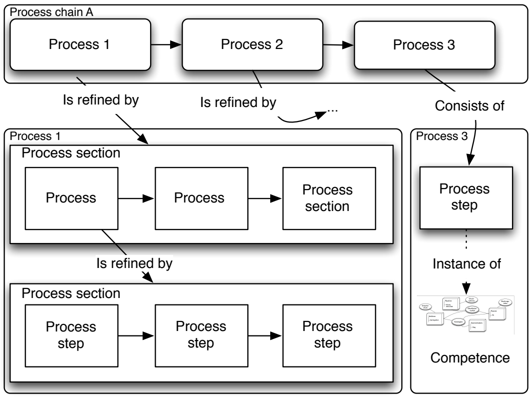


Figure 2. Hierarchical process chains for modularization

A. Hierarchical Process Chains and Variants

In process modeling for *MinaBASE* it is crucial to arrange process chains hierarchically for reducing complexity. In this way, basic information can generally be presented on higher levels, while details are hidden on lower levels. This is illustrated in Figure 2.

The subelements of process chains are single processes. Hierarchical process chains are characterized by the fact that the process elements involved can be refined in any way. In the example the first process is refined by a so-called process section. Process sections are a special process element, they represent well-defined workflows, such as the LIGA process in microsystems technology, which combines the techniques of lithography, electrodeposition, and molding [6]. The first element of the upper process section is refined by another process section which consists of atomic process steps only. They correspond to an instancing of competences. If this direct allocation of process step to competence does not yet exist, e.g., in the early development stage of the process chain, it must be possible to model technologies within an atomic process element as well as complete subprocesses as alternative variants for a part of the process chain.

B. Integration of Product Structure and Process Chains

To describe the setup of a microsystem, it is recommended to store its components in a product structure (cf. Figure 3), by means of which a microsystem can be set up and structured according to the construction kit principle. The first structuring means are modules that encapsulate a logical functional area and hide the details from other modules. Individual components having certain functions can be joined to a logical entity by assemblies. Theoretically, a module is composed of a few single components or of a large number of assemblies. To produce a complete microsystem, the process chains resulting in components and the integration of these components are relevant. Joining of the individual components and assemblies requires or represents an own type of processes integrating the components as results of the process chains of the microsystem.

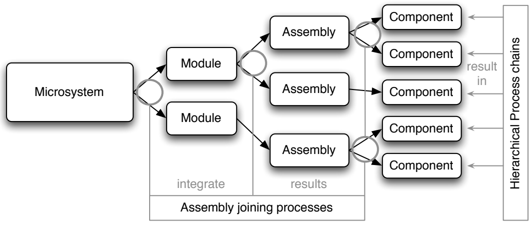


Figure 3. Allocation of process types to the product structure

C. Versioned Documentation of Solution Approaches

On the way towards finding a solution for a technical problem, many data are generated, which will be very valuable when solving future problems [3]. If acquired at all, they exist, e.g., in the form of a text or table document often decentrally without any relationship to the solution process. Therefore, these data have to be combined manually. This is aggravated by the fact that the data are often administrated by different persons and exist in a heterogeneous structure. Another obstacle lies in the fact that success only is documented in many cases, but not the errors made on the way towards it. But it is the information whether and why a certain approach did not work in the past, which needs to be available when implementing new ideas, such that errors already made will not be repeated. These circumstances frequently make the implementation of new ideas time-consuming and expensive. Good ideas are rejected in the beginning already, only because an adequate and collaborative access to information is lacking [7]. To overcome this problem, process chains must be stored in various versions for the same technical problem or components of the product structure and easy to reproduce. A new version may result from the fact that experiments for the process model applied so far have shown that parameters of certain materials affect the quality required in an unexpected adverse way.

D. Derivation of Templates from Ad hoc Workflows

A decisive criterion of efficient knowledge management is access and an efficient reusability of knowledge entities. In microsytems technology some basic processes are distinguished, such as ”disposable” or ”AVT”. They require similar workflows in each case. It must therefore be possible to derive project-independent patterns from project-specific process chains, which may then serve as documentation or templates for new projects by individual instancing, as shown in Figure 4. Patterns are supposed to accelerate the development of new products by reusing existing templates.

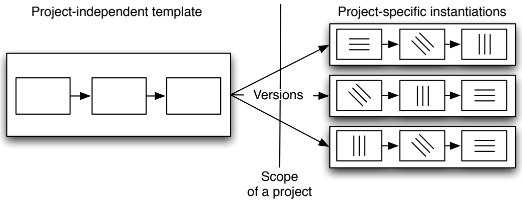


Figure 4. Abstraction of templates by project-specific instancing

E. Modeling of Dependences

Apart from defining the structure and order of process elements, it is important to express dependences of process elements. This significantly extends the capability of modeling process chains, as not only the structure alone is of interest, but also the fact why exactly this structure is required. In addition, process elements may have certain TP as a prerequisite. The possibility of expressing the following types of dependences explicitly facilitates the finding of errors in an early phase, in which these errors can be eliminated at low costs.

- *PreConditions* - A process element is executed under a condition, for example, the existence of an applied layer.
- *DuringConditions* - Circumstances prevailing while executing a process element, for example, the temperature of a production process. When using a process, TP may occur, which can now be documented.
- *PostConditions* - Express that consistency conditions, such as the observation of a certain TP, shall remain valid even after the execution of the process element for all following process elements.
- *Effects* - Are the result of a process element, for example, reaching of property required by a specification. Previous post-conditions are overwritten, if, e.g., a lacking thermal resistance due to the application of an insulation layer is defined as an effect of a later process element.

Having formulated own conditions and effects, other technical experts can compare them with own dependences so that cooperation is supported. Modeling of the dependences provides for an explicitly formulated representation that can be communicated and evaluated automatically.

IV. MinaBASE Process Modeling

This section will focus on standards of business process modeling. Then, their suitability for *MinaBASE* process modeling will be studied taking into account the requirements described above. Finally, the *MinaBASE* process model designed will be presented.

A. Standards of Process Modeling

1) Event-driven Process Chains: Event-driven process chains [8] are part of the ARIS concept (architecture of integrated information systems) [9], in which they act as graphical, semi-formal modeling language for business processes. An event-driven process chain is a directed graph, whose nodes consist of alternating events and functions as well as logical connectors. An event is a state initiating a function or initiated by the latter. Logical connectors allow for the splitting and combination of parallel or alternative workflows by the interconnection symbols `AND`, `OR`, and `XOR`. The ”process path” is used to reference partial processes. By ”extended event-driven process chains”, event-driven process chains are extended by notations for organizational units, data objects, and services.

2) Petri-Nets: Petri nets [10] have a formal mathematical basis [11], [12] and are used for modeling and simulating business processes as well as for conceiving concurrent and parallel algorithms. A Petri net is a directed, bipartite graph containing two types of nodes, the transition for events and the place for conditions. Places may contain marks. During the so-called ”firing”, these marks are removed from the input places of a transition and newly generated in the output places, as a result of which the net can be run as a simulation. Hierarchical Petri nets [13] allow for the storage of partial processes in own nets. The predicate transition nets contain structured marks representing objects [14], whose state can be modified by calculations in transitions.

3) Business Process Modeling Notation: The Business Process Modeling Notation (BPMN) [15] is a semi-formal modeling language for business processes with a small formal basis. BPMN is used to define a workflow that is translated into languages like BPEL (Business Process Execution Language) in order to integrate web services in the processes by a ”service-oriented architecture”, for instance. A central element is the ”business process diagram” that consists of ”flow objects” (atomic and composed activities for subprocesses, gateways as connectors, notations for events), ”connecting objects” (edges for the workflow and information flows), ”swimlanes” (allocation of roles), and ”artifacts” (information objects and metadata) [16].

4) Activity Diagrams: The Unified Modeling Language (UML) is a formal and visual modeling language for the design and documentation of artifacts of software systems. The UML defines various types of diagrams [17] to model, e.g., the structure (class and component diagrams) and behavior (sequence and activity diagrams). Activity diagrams were used for the process specification language [18] to model production processes. They contain actions as elementary elements that model complex behavior by chaining with control and object flows and logical connectors. Control flows are directed edges specifying the sequence of actions. Object flows extend this semantics to represent data flows of objects along an edge. An action of an activity can be structured hierarchically to represent the exact workflow in another diagram.

B. Applicability of the Process Modeling Standards

The first requirement in Section III-A deals with the hierarchization of process chains and the possibility of representing variants. All standards analyzed in IV-A support own forms of subprocesses that can be referenced. However, they do not directly provide for the modeling of technical variants. It is possible to note several variants in a running text indexing symbols for functions, but this is a rather informal approach that is difficult to evaluate automatically when verifying the process model in terms of the dependencies modeled. For this reason, the requirement is met partly only. Requirement III-B asks for an integration of process chains with the components of a microsystem to make it clear which component is produced by which process chain. *ARIS* defines a performance view for the representation of results of process chains, a general product model, a hierarchical product tree for the event-driven process chains meeting the requirement. Petri nets do not support any form of product models and do not meet the requirement. BPMN and activity diagrams allow for the modeling of unstructured objects. Activity diagrams support typed object flows. BMPN uses ”data objects” that stand for used documents. As both standards do not allow for hierarchical objects, they meet the requirement partly only. Requirement III-C focuses on the support of an iterative, collaborative, and centrally available project documentation of solution approaches for them to be used for the development of new microsystems and for the later reproduction of errors and experience gained during previous developments. As all standards support a serialization by, e.g., XML data formats, this requirement can be met in principle by all standards. Requirement III-D covers the storage of process chains as templates for new processes. In principle, all standards are capable of using so-called reference processes given in a top-down manner. They have to be adapted to the existing conditions. As a derivation of process templates from existing processes is much more important for *MinaBASE*, however, the standards meet this requirement partly only. Requirement III-E deals with the modeling of dependences of process elements. Due to the informal character of event-driven process chains, dependences can be expressed as running text only, such that the standards do not meet the requirement. In Petri nets events are modeled, which are executed only after their preconditions are met and result in post-conditions. These are expressed by the structure of the process chain. This aggravates the allocation of knowledge entities of *MinaBASE*, such as TA and TP or effects for product properties as functions of individual process elements. Consequently, Petri nets meet this requirement partly only. The UML contains the Object Constraint Language (OCR), such that constraints and conditions can be modeled. However, programming knowledge in OCR is required. BPMN has explicit language constructs for the waiting for the receipt of messages or other events, such that a simple type of dependences can be modeled. Consequently, BPMN and activity diagrams meet the requirement partly.

To sum up, none of the standards described fulfils all requirements to the complete extent. As a result of the ARIS concept, event-driven process chains are the standard fulfilling most of the requirements, with informal modeling aggravating the storage of dependences that can be evaluated automatically. BPMN lacks the formal semantics of a metamodel for being extended such that the limits of graphical notation are overcome. Petri nets do not fulfill two requirements and in spite of their strongly formal basis, they can hardly be applied for *MinaBASE* process modeling. It is found that in spite of varying strengths and weaknesses, none of the standards fulfils the requirements to a sufficient extent and that an individually tailored knowledge representation is suited best.

V. Perspective Model for *MinaBASE*

A generic metamodel for standards of process modeling was developed in [19]. Based on this model, the standards described above (see Section IV-A) were evaluated. When comparing this metamodel with the requirements made on *MinaBASE* process modeling, it is found that even the metamodel that combines the modeling capabilities of many standards by various perspectives meets the requirements to a limited extent only. Hence, it is absolutely necessary to conceive an individually tailored solution, since an expressive methodology for business process modeling is not suited for *MinaBASE*, because the complexity does not lie in the branching of activities, but in the information, concepts, their relationships and constraints associated with the activities, and above all in the implicit knowledge resulting from combination. An integrated solution approach to knowledge representation has to define own perspectives in order to close the gap between the requirements and the generic metamodel and to delete the unused perspectives from the metamodel. The multi-perspective model developed here is shown in Figure 5. This model is divided into four overlapping and integrated perspectives and, hence, facilitates the management of complexity and reuse of individual knowledge entities. The system perspective contains the product structure from requirement III-B, i.e. the structured grouping of microsystems and the specifications to be fulfilled by the components. Consequently, this perspective covers everything relating to the setup and functions of a microsystem. The process perspective covers the requirements from III-A, i.e. modeling of workflows in the process chain and the hierarchical structure and variants of possible technologies and competences. The link between the process and system perspectives is the allocation of components of the product structure as results of process chains from requirement III-B. The development perspective expresses the requirement III-E, i.e. modeling of pre-, during-, and post-conditions and of effects of process elements. This separates the logical setup of process chains from their partly automatic verification and validation on the basis of the constraints and dependences expressed by conditions. The project perspective covers the requirements III-C and III-D. Provided that *MinaBASE* is used properly, it is possible to view versions of the process chains and product structures in a historically reproducible manner. This results in an iterative product development cycle and may reduce the consumption of resources of future developments in terms of time and costs.

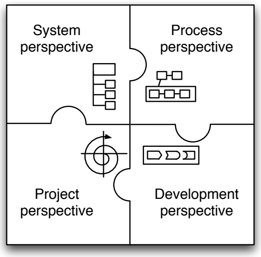


Figure 5. Model for *MinaBASE* process modeling

VI. Modeling in OWL

It is the objective of *MinaBASE* to acquire process knowledge in microsystems technology such that the finding, combination, and, ideally, verification of knowledge can be supported automatically. To ensure this for the model described, it is reasonable to model the perspectives and their concepts by formal knowledge representation languages from the semantic web, e.g., OWL (Web Ontology Language) [20], as this makes the semantic relations described explicit. After acquiring the process knowledge in this form, the next step may be a definition of rules, e.g., with SWRL (Semantic Web Rule Language) [21] and queries of increased content values using SPARQL (SPARQL Protocol and RDF Query Language) [22]. Previous approaches, such as the ARIS concept [23], define a semantic model, but the significance of the contents of the process elements is lost rapidly by marking with a purely free text. An adequate support of the modeler in semantic annotation, i.e. the filling of the process models with contents, usually does not take place [24], such that hardly any automation and machine support is possible, since the interpretation of the concepts, terms, and relations used is left to a human brain. For this reason, knowledge representation of the multi-perspective model is accomplished by formal languages from the semantic web. As it is a principle of *MinaBASE* to separate the so-called build-time from the runtime, i.e. to define the concepts used to structure the production knowledge during runtime, however, the build-time must have a flexible structure allowing for later instancing and adaptation. Hence, the

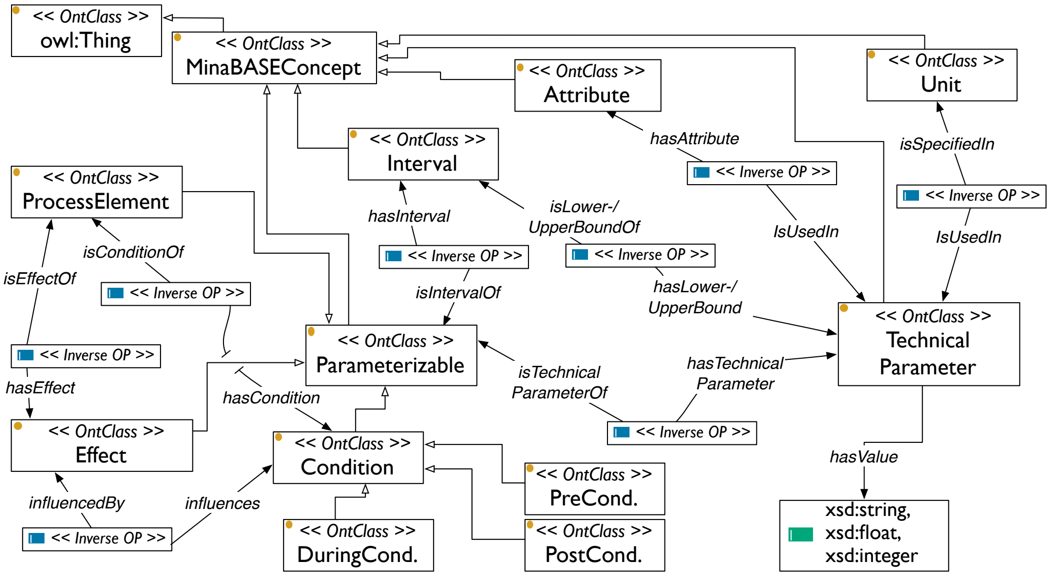


Figure 6. Schematic representation of classes and relationships of the development perspective in OWL

OWL ontology to be generated for the model perspectives should rather be an ”upper-level ontology” allowing for the instancing and allocation of process knowledge during runtime. Figure 6 displays a schematic representation of the modeling of the development perspective using the OWL language, which will now be explained in more detail. As a graphical notation, the visual metaphors of the OWL editor *Protege* are used, with a circle representing a class and a rectangle an ”object” or ”data type property”. The uppermost element of the class hierarchy is the class *MinaBASEConcept* that inherits from *owl:Thing* and serves as a central extension point for the classes of the respective perspectives. While the process perspective contains the hierarchical setup of the process elements, the development perspective models the dependences in the process chain using TP, conditions, and effects. The class *ProcessElement* possesses general relations for input and output edges to other process elements in the process perspective and acts as basic class for composed *CompositeProcessElements* like *ProcessSection* and atomic process elements like *ProcessStep*. Hence, the perspectives can be linked without having to combine the details of the perspectives. A central class of this perspective is *Parameterizable* that encapsulates the allocation of TP by the inverse ObjectProperty *hasTechnicalParameter*, such that subclasses of Parameterizable, namely, ProcessElement, *Effect*, and *Condition*, inherit this relation. The classes of *PreCondition*, *DuringCondition*, and *PostCondition* are derived from *Condition*. Their existence is required for referencing in rule languages and implementing their semantics. An effect is linked with conditions via the *influences* relation. In this way, preconditions of previous process elements can be relaxed. By *hasCondition* or *hasEffect*, instances of these classes are allocated to the process elements. Apart from the TP, also intervals can be allocated to instances of the class *P*arameterizable as value ranges by *hasInterval*. The class *Interval* may contain upper and lower limits of TP using the relations *hasLowerBound* and *hasUpperBound*. A TP is implemented by the class *TechnicalParameter* and possesses a data type property *hasValue* for typed values in the form of integers, floating-point numbers or character strings. These are specified in a certain unit via the class *Unit* and reference an *Attribute*, for example, the density. Using the inverse relation *isUsedIn* of the *hasAttribute*, TP comparable in rules can be determined in order to determine proposals for the allocation of new TP when reusing existing process elements for the modeling of new process chains.

VII. Conclusion and Future Work

This paper presented an approach to extending the *MinaBASE* process knowledge database, a system for managing the knowledge in the field of microsystems technology. By means of this approach, the knowledge entities of the system, basic data on processes, materials, and production competencies, can be combined in a process-oriented manner. First, the modeling requirements were presented, which result from the special characteristics of microsystems technology compared to conventional mechanical engineering. In particular, the interdisciplinarity of the expert knowledge required, the low standardization of production methods, and the necessity of an iterative solution of development problems characterize microsystems technology. Central requirements in *MinaBASE* are process chains that can be interlinked in many ways along an "is-Part-of" hierarchy, the modeling of product structures and their allocation to process chains, the definition of dependencies within the process chain, a versioned storage of project-specific documentation, and the possibility of deriving project-independent and reusable process templates for new process chains. Then, standards of process modeling (event-driven process chains, Petri nets, BPMN, and activity diagrams of the UML) were analyzed for their suitability for meeting the requirements described. As none of the existing standards meets all requirements, an individual, multi-perspective, and interlinked model was conceived, which is tailored to meeting these requirements. To not only store the production knowledge in *MinaBASE*, but derive new knowledge from implicit relationships within the knowledge base, technologies from the semantic web were selected to implement the model conceived. As a first step, the concepts and relations of the model were formulated as ontology in OWL and the development perspective was presented. Extension of the ontology by rules for the implementation of the semantics of conditions and effects of the process elements, integration in the existing application architecture of *MinaBASE*, testing of the ontology using data from practice, and a possibly resulting refinement of the concepts selected and relations have been identified as future research topics.

REFERENCES

[1] I. Nonaka and H. Takeuchi, "The knowledge-creating company," *Harvard Business Review*, vol. 6, pp. 96–104, 1991.

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

[3] U. Hansen, C. Germer, S. Büttgenbach, and H. Franke, "Rule based validation of processing sequences," in *Techn. Proc. MSM*, 2002.

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

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

[6] E. Becker and W. Ehrfeld, "Das LIGA-Verfahren–Mikrofertigung durch Röntgentiefenlithographie mit Synchrotronstrahlung, Galvanoformung und Kunststoffabformung," *Phys. Bl*, vol. 44, no. 6, pp. 166–170, 1988.

[7] D. Ortloff, J. Popp, K. Hahn, T. Schmidt, and R. Bruck, "Tool Support for Microelectronic Process Development," *Mixed Design of Integrated Circuits and Systems, MIXDES 2008*, pp. 467–472, 2008.

[8] A.-W. Scheer, "Semantische Prozeßmodellierung auf der Grundlage Ereignisgesteuerter Prozeßketten (EPK)," *Veröffentlichungen des Instituts für Wirtschaftsinformatik*, 1992.

[9] ——, "ARIS-House of Business Engineering," *Veröffentlichungen des Instituts für Wirtschaftsinformatik*, vol. 133, 1996.

[10] C. A. Petri, "Kommunikation mit Automaten," Ph.D. dissertation, Institut für instrumentelle Mathematik, Bonn, 1962.

[11] T. Murata, "Petri nets: Properties, analysis and applications," *Proceedings of the IEEE*, vol. 77, no. 4, pp. 541–580, 1989.

[12] J. E. Jorg Desel, *Free Choice Petri Nets*. Cambridge University Press, 2005.

[13] R. Fehling, *Hierarchische Petrinetze: Idee und grundlegende Struktur.* Universität Dortmund, Germany; Lehrstuhl Informatik 1, Forschungsbericht Nr. 344, 1990.

[14] P. Elgass, H. Krcmar, and A. Oberweis, "Von der informalen zur formalen Geschäftsprozeßmodellierung," *Geschäftsprozeßmodellierung und Workflow-Management. Modelle, Methoden und Werkzeuge. Bonn, Albany: Internat. Thomson Publ*, pp. 125–139, 1996.

[15] S. White, "Introduction to BPMN," *IBM Cooperation*, 2004.

[16] OMG, "BPMN 1.2 - Final Adopted Specification," 2009.

[17] ——, "Unified Modeling Language, Superstructure 2.2," 2009.

[18] C. Schlenoff, M. Gruninger, T. Creek, M. Ciocoiu, and J. Lee, "The Essence of the Process Specification Language," *Transactions of the Society for Computer Simulation*, vol. 16, pp. 204–16, 1999.

[19] B. List and B. Korherr, "An evaluation of conceptual business process modelling languages," in *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2006, pp. 1532–1539.

[20] M. Dean and G. Schreiber, "OWL Web Ontology Language Reference," W3C, W3C Recommendation, 2004.

[21] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML," World Wide Web Consortium, W3C Member Submission, 2004. [Online]. Available: http://www.w3.org/Submission/SWRL

[22] A. Seaborne and E. Prud'hommeaux, "SPARQL Query Language for RDF," *W3C Recommendation*, 2008.

[23] C. Fillies and F. Weichhardt, "On Ontology-based Event-driven Process Chains," in *GI-Workshop EPK*, 2005.

[24] B. Heinrich, M. Bewernik, M. Henneberger, A. Krammer, and F. Lautenbacher, "SEMPA–Ein Ansatz des Semantischen Prozessmanagements zur Planung von Prozessmodellen," *Wirtschaftsinformatik*, vol. 50, no. 6, pp. 445–460, 2008.

A Three-Tier Matching Strategy for Predesign Schema Elements

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

Jürgen Vöhringer
Institute for Applied Informatics
Alpen-Adria Universität Klagenfurt
Klagenfurt, Austria
e-mail: juergen.voehringer@ifit.uni-klu.ac.at

***Abstract*—Schema integration is a very complex task in which several schemata are merged into one global conceptual schema. Due to its complexity, computer-based applications and tools are needed that support and automate parts of the integration process. In our previous work, we have shown that schema integration on the predesign level allows for lower schema complexity, fewer conflicts and better end user feedback. In this paper, we focus on applied matching strategies, which are a central element of any semi-automatic integration process. We propose a set of matching methods that are suitable for the predesign level and discuss how they are intertwined and how their results regulate the integration process. As its main contribution, the paper offers an integration of previously presented methods and describes exemplary findings from the corresponding prototype.**

Keywords- predesign modeling; matching; integration

I. INTRODUCTION

When designing and developing information systems, we often have to deal with requirements, hereafter referred to as schemata, which are collected from different sources. These requirements are often divided into structural and behavioral schemata. In this paper, we focus on the structural aspect, meaning both what data should be stored in the database and what data the information system needs for processing its functionality. The application area is schema integration, a very complex task in which several conceptual schemata are merged into one global conceptual schema. In [3], the authors define schema integration as “the activity of integrating the schemas of existing or proposed databases into a global, unified schema” (p. 323). Due to complexity, computer-based applications and tools are needed in the integration process to help users not only to recognize but also to resolve similarities and differences between two source schemata.

In this paper, we mainly focus on the former of these: the recognition of similarities and differences. In doing so, we propose a three-tier matching strategy for predesign schema elements starting with an element level matching approach followed by structural level matching and ending with a taxonomy-based matching strategy. Our approach is rather unique since it is modeling-language independent, which means that the matched schemata can be formulated in any modeling language focusing on concepts and links between concepts. In our approach we focus on linguistic techniques to extract as much information as possible. Matching on the predesign level has various application areas, including

A. Integration of heterogeneous requirements during the early phases of information system development projects,

B. Consolidation of project schemata from a specific domain during ontology engineering.

Generally speaking, predesign matching is best used whenever natural language descriptions are available rather than more formal specifications (e.g., during requirements engineering), when semantic, project-overarching matching is needed (e.g., during ontology engineering) or when extensive user feedback by domain experts is expected or required [10][1]. In [26], we describe an experimental study that compared end user feedback for predesign models compared to feedback for standard conceptual models such as Unified Modeling Language class diagrams.

The paper is structured as follows: in section two, we address some related work and distinguish it from our own approach. In section three, we present our three-tier matching strategy consisting of element level matching, structural level matching and taxonomy-based matching. In section four, we show how these three tiers are interconnected and how the matching results are utilized for the purpose of schema integration. Finally, the paper closes with a summary and conclusions.

II. PREVIOUS AND RELATED WORK

Earlier work in the domain of schema integration might be roughly classified into three approaches [4]: manual, formal and semi-automatic approaches to schema integration. *Manual* means that all tasks are performed by hand, *formal* means in this context that a formal modeling language is applied and *semi-automatic* means that at least one computer-based application is used in the integration process. Looking at previous work, it can be concluded that much of that work has focused on the Entity-Relationship modeling language (ERML) [12] or some extension of it [25]. Lately, focus has shifted towards the Unified Modeling Language (UML) [20]. Even so, it should be noted that both the ERML and the UML are implementation-dependent modeling languages. In our approach, we instead work towards a method for modeling-language independent schema integration, meaning focus is on content instead of implementation. In the rest of this section we give examples of semi-automatic approaches and distinguish them from our own approach.

In [23], the authors present a survey of approaches to automatic schema matching. They distinguish schema-based and instance-based matching. Our work is classified as a schema-based approach, since it is applied early in the information system development process in which schemata are focused. In [23], the authors further state that schema-based matching can be performed on the element level (concept) and on the structural level (neighborhood) and it can be either linguistic or constraint-based matching. Our approach is a composite schema-based matching approach since we apply element level matching, structural level matching and taxonomy-based matching. The work in [23] was also adapted and refined in [24].

In [17], the authors present a method for structural conflict resolution while applying the ERML. The authors pinpoint that in their method, structural conflicts are automatically resolved resulting in less manual effort. Finally, in [17], the authors state that "the key structural conflict is that between an entity type and an attribute" (p. 227). In our approach, we do not distinguish between entities (classes) and attributes because we work on a higher level of abstraction compared with the ERML and the UML.

In [14], the authors once again adopt the ERML while addressing schematic discrepancy. The authors present algorithms that resolve the problem. In the algorithms, meta-data are transformed into entities and the authors pinpoint that the information and constraints given in the source schemata are kept. Similar to the work presented in [17], the work presented in [14] is classified as work on an implementation-dependent level.

Several algorithms for calculating concept similarity have also been proposed such as the Wu and Palmer metric [30], the Hirst and St Onge metric [15] and the Lesk metric [2]. All three algorithms will be presented in more detail in section 3C, taxonomy-based matching.

Finally, we have found that some techniques of our matching strategy are similar to the ones used in the DIKE approach [21] and the GeRoMeSuite [16]. However, the DIKE approach is quite different compared to ours since in our approach we do not focus on any specific modeling language but instead only on concepts and links between concepts. In the GeRoMESuite [16] the authors present a system for holistic generic model management but their approach focuses on implementation dependent models such as SQL, XML and OWL, while our focus is on implementation independent conceptual schemata, meaning the approaches are complementary rather than exclusively.

III. Matching Strategies

An important aspect of our semi-automatic three-tier matching approach is its independence from any specific modeling language [9], meaning it can be used for the integration of schemata that are available in different source languages. Of course, this also means that our strategy cannot depend on language specific modeling concepts but has to utilize other, e.g., linguistic, information to analyze the models.

In our approach we first perform comparisons on the *element level* for gathering preliminary matching proposals. Then *structural level matching* is applied to identify potential contradictions to the original assumptions that might hint at homonym or synonym conflicts. Finally, we use an optional *taxonomy-based approach* to identify previously undetected concept relationships. The latter step is especially relevant when concepts are matched in the context of ontology engineering, since it has the potential of detecting hidden, easily overlooked information.

All of these strategies are applied on concept pairs with both members of the pair coming from one of the matched schemata. Thus, in preparation for the matching process, all relevant concept pair permutations are generated – since the pairs are symmetric, the order of the concepts in the pair is irrelevant. In a further preprocessing step, linguistic tools like stemmers and lemmatizers are used to reduce the words from the concept designations in the target schemata to their base forms [8].

The following sections will describe the different levels of the matching approach in more detail.

A. *Element level matching*

On the element level, concepts are directly compared to each other without considering the context. The main matching criteria on this level are the names of concepts; element level matching therefore presupposes that the concept names are available in their linguistic base form. Other matching criteria on the element level are potentially available metadata such as definitions, indications of quantity or data types, though the latter is implementation-dependent and thus typically not available on the predesign level.

The eventual goal of element level matching is to decide whether a concept pair matches. The process has the following possible outcomes:

- Equivalence/Synonymy,
- Relatedness,
- Independence.

At first glance, equal words/definitions suggests *equivalence*, although the concepts might be later identified as homonymous. If the compared concept names are not equivalent, but domain ontology is available and both compared words describe known ontology concepts, then information about the *relatedness* of the compared concepts is queried. If they are not synonymous they may be directly related in another way, indirectly related via several intermediate concepts or completely unrelated. If potential relatedness is detected in the ontology, this information is incorporated in the integration proposal.

If the compared concepts are classified as *independent* after the first matching steps (i.e., no potential relatedness was found), but one or both of their names consist of compound words, then these names are deconstructed. For endocentric compounds – the most common ones in the English language [11] – the right-most element of the compound word is its head. Thus, the following percolative rules are applicable for identifying automatic relationships between the words:

A. If the compared concept names are available in the form of A and AB (i.e., A corresponds to the compound AB minus the head B), then the relationship "AB belongs/related to A" can be assumed.

B. If the compared concept names are available in the form B and AB, where A is the head of the compound AB, then the relationship "AB is a B" can be assumed.

To exemplify the first rule, the concept "car color" is identified as a potential attribute of "car" ("car color" belongs to "car"), and the concept "student name" is identified as an attribute candidate for "student" ("student name" belongs to "student"). Regarding the second rule, the exemplary concept "dialysis patient" would be interpreted as a "patient" ("dialysis patient is a patient"), and "blood pressure measurement" is a (specific form of) "measurement".

On a related note, if no definition or ontology data is available about a schema concept, semi-automatic disambiguation can be attempted, using generic lexicons such as WordNet [19] that contain word sense definitions. The word in question is looked up in the lexicon, which results in all possible word senses and their definitions being returned. The following four outcomes are possible:

- exactly one definition is returned;
- more than one definition is returned;
- no fitting definition are returned;
- the returned definition is on the wrong detail level.

If more than one meaning is returned, the senses are ranked according to their likelihood of occurrence in the English language or the domain in question. If no meaning is returned, other searches are automatically attempted with linguistic decompositions or variants of the word. If the returned definition is on the right track, but on the wrong detail level, the search is repeated for the candidate concept's hypernyms or hyponyms respectively. The entire process is described in detail in [28].

B. *Structure level matching*

On the structural level, comparisons of the concepts' neighborhoods are conducted, meaning that those concepts that are directly connected to concepts, which have been identified as equivalent or similar to concepts in another source schema, are compared. In doing so, several similarities and differences might be recognized that otherwise could pass unnoticed. Besides that, structure level matching is also used to verify or decline the results of element level matching. In structure level matching, we propose to use a set of "IF THEN" rules. Moreover, certain influence factors such as *polysemy count*, *valency* and *domain weight* might be used to complement the rules. The influence factors could even be used to decide whether neighborhood comparison is necessary [8] at all.

Polysemy count gives the number of meanings a word has in a given language, valency gives the number of attribute slots a word has in a given language and domain weights can be manually assigned to concepts by domain experts [8].

We propose to use two types of "IF THEN" rules: *rules for equivalent concept names* and *rules for similar concept names* (see also [6][7]). As the rule names indicate, *equivalent* means that two concept names are recognized as equivalent in element level matching, e.g., "Name" in schema 1 and "Name" in schema 2. *Similar* means that the concept names are not equivalent but recognized as similar in element level matching, e.g., "Order" in schema 1 and "OrderItem" in schema 2. In total, at least six rules should be used for equivalent concept names and three rules for similar concept names. *The rules for equivalent concept names* can be stated as:

IF comparison of concept names yields equivalent and comparison of concept neighborhoods yields:

- Equivalent THEN *equivalent* concepts are most likely recognized.
- Different THEN *homonyms* are most likely recognized.
- Similar AND one concept in each schema is named different, THEN *synonyms* are most likely recognized.
- Similar AND one concept name is a composite of another concept name with a following addition, AND cardinality is indicating 1:1, THEN an *association* between the two concepts is most likely recognized.
- Similar AND one concept name is a composite of another concept name with a prior addition, THEN a *hypernym-hyponym* pair is most likely recognized.
- Similar AND one concept name is a composite of another concept name with a following addition AND cardinality is indicating 1:M with or without uniqueness, THEN a *holonym-meronym* pair is most likely recognized.

The rules for equivalent concept names verify or decline the result from the first part of element level matching, while the rules for similar concept names verify or decline the result from the second part in which the percolative rules are applied. The *rules for similar concept names* can be stated as:

IF comparison of concept names yields:

- Similar, one concept name is a composite of another concept name with a following addition, AND comparison of concept neighborhoods yields similar or equivalent with an indication to a 1:1 cardinality THEN an *association* between the two concepts is most likely recognized.

- Similar, one concept name is a composite of another concept name with a following addition, AND comparison of concept neighborhoods yields similar or equivalent with or without an indication to a unique 1:M cardinality THEN a *holonym-meronym* (aggregation) pair is most likely recognized.
- Similar, one concept name is a composite of another concept name with a prior addition, AND comparison of concept neighborhoods yields similar or equivalent THEN a *hypernym-hyponym* pair is most likely recognized.

In [6] and [7] it was described how the rules could be applied while applying the Karlstad Enterprise Modeling approach [13]. However, in this paper we do not focus on any specific modeling language and therefore we have also refined and adapted the rules to be useful for any modeling language; in other words the rules are modeling-language independent.

C. Taxonomy-based matching

The previous matching strategies for concept pairs were all based on their names and context or the use of domain ontologies. Domain ontologies, however, are not always available, and concepts may have a sparse neighborhood, which can make analysis of their context unreliable. Using general-purpose taxonomies that are not restricted to one domain, general assumptions about the relationship between two words can be made: isolated words are compared based on their position in the taxonomy instead of on their structure or context.

A particularly extensive domain-independent taxonomy for the English language is provided by the lexical database WordNet [19], which is freely available and thus widely used in scientific research projects. It is important to note that using WordNet for calculating concept similarity is completely separate from using WordNet for disambiguation purposes as discussed on the element level. In [18], the authors compared a number of different approaches for calculating semantic similarity metrics based on WordNet. Perl-based implementations for deriving concept similarity measures from WordNet were also presented in [22]. Among them, Wu and Palmer, Hirst and St Onge and Lesk will be shortly discussed here because they are three very different forms of WordNet-based similarity measures.

The *Wu and Palmer metric* was first suggested in [30]. The similarity value is calculated using formula 1:

$$wup = \frac{2 * depth\ (LCS\ (concept\ 1, concept\ 2))}{depth\ (concept\ 1) + depth\ (concept\ 2)} \qquad (1)$$

In a first step, the least common subsumer (LCS) is determined, i.e., the first common parent of the compared concepts in the taxonomy. The similarity score is derived from dividing the double of the taxonomy depth of the LCS (since two concepts are compared) by the sum of the taxonomy depths of the compared concepts. Further separation of the concepts from their first common father concept means a lower similarity score.

The *Hirst and St Onge metric* [15] allows measuring the similarity between two concepts by determining the length of the taxonomy path between them. Three different kinds of paths for connecting concepts can be distinguished based on their strength: *extra-strong*, *strong* and *medium* paths. Extra-strong paths exist between two equivalent concepts. Strong path are identified by a direct connection between two concepts. Medium-strong paths finally mean that two concepts are indirectly connected. In the latter case, the number of path direction changes is relevant for determining the concept similarity. Direction changes occur every time a medium-strong connection switches between upwards-paths (generalizations), downward-paths (specializations) and horizontal paths (other relationships between concepts). Frequent direction changes lower the similarity score, as shown by formula 2:

$$hso = C - pathLength - k * numberDirectionChanges \qquad (2)$$

The calculation returns zero if no path at all exists between the concepts. In that case, the concepts are interpreted as unrelated. C and k are constants used for scaling the metric.

Finally, the *Lesk metric* [2] is a context-based similarity score that does not require taxonomic structures. Instead it presupposes a lexicon, in which different word senses are distinguished and detailed definitions for each meaning are available. Because the WordNet taxonomy contains definitions and examples for each concept, it is a popular choice for this task. For determining Lesk similarity, the definitions of both involved concepts must be provided; then a numerical estimation of their degree of separation is calculated by counting the word overlap.

Traditionally, the Lesk algorithm is used for disambiguating words in full natural language texts: a context window containing an equal number of words on both sides of the observed word is defined. Then all available definitions for the observed concept and the other content words in the context window are examined and compared, ignoring non-content words such as pronouns or articles. The word sense that has the greatest overlap with the definitions from the surrounding text is assumed to be the correct one.

In our use of the Lesk algorithm, already disambiguated concept-pairs are presupposed and the Lesk metric is used to calculate similarity scores for them. The scores describe the concept pair's relative similarity compared to other concept pairs and – if an according threshold value has been defined – the conflict potential of the word pair. For example, using our optimized Lesk algorithm, the concept pair "car"-"bicycle" has a similarity score of 198, "car"-"motorcycle" has a score of 321 and "car"-"bus" is assigned the score 688, which indicates their relative similarity.

The algorithm and potential optimizations of the Lesk algorithm for our purposes was described in detail in [28]. Lesk is the most relevant WordNet similarity measure for the matching purpose since it is rather robust against inadequacies in the taxonomic structure and its results can be improved by relatively simple, light-weight enhancements of the taxonomy, such as filling gaps in concept definitions.

The results of taxonomy-based concept matching are a starting point for future ontology extensions. If, for instance, a high matching score is identified between two previously unrelated concepts, then a relationship between them can be assumed, which is a candidate to be incorporated in the domain ontology.

IV. Further Use of the Matching Results

Any integration attempt needs to follow a predetermined workflow that combines the various techniques that have previously been described. In [10], this process was called Concept Determination Method (CDM), because at the heart of it is the disambiguation of concepts, which is a precondition for conflict recognition and resolution, which in turn enables model consolidation. The following parties are involved in the process: system designers, domain experts and ontology supervisors. The process input typically consists of two schemata, which are to be integrated. A single schema is also a permitted input; in this case only the schema-preprocessing phase is traversed, which involves the optimization of its modeling element designations and the resolution of any potential inner-schema conflicts. In all cases, the output of the CDM consists of one single (integrated) schema. The integration process is supported by a number of external repositories, namely the domain ontology and an optional domain-overarching taxonomy/lexicon. For the purpose of testing the CDM, an integration prototype is currently under development at Alpen-Adria-Universität Klagenfurt.

The CDM starts by choosing the source schemata that should be consolidated, one of which is typically the current integrated schema. If more than two schemata need to be integrated, they are processed pair-wise one after the other, with the current integrated schema always being one member of the pair. In cases where only one schema is chosen as input, it is preprocessed and returned in optimized form. The integration process itself follows the typical phases (1) schema preprocessing, (2) schema matching and (3) schema consolidation.

In preprocessing, schemata are examined for internal conflicts and prepared for the following phases. Afterwards, the matching phase begins, which consists itself of several sub phases that were described earlier in the article. How the several matching techniques are utilized in a common workflow was first discussed in detail in [8].

In the first step of schema matching, all permutations of concept pairs from the two source schemata are prepared for comparison. The eventual matching goal is to decide whether the compared concepts are the same or different. The proposed workflow is as follows: every concept pair is first matched on the element level using the direct comparison of the base form and the application of linguistic rules. This step results in a preliminary matching decision. If the result is "independent" and a domain ontology is available, then information about potential connections between the concepts are looked up in the ontology. Technically speaking, this is still a part of element level matching, because the concepts' context is irrelevant for this step.

Concept pairs that have been classified as "independent" or "equivalent" during element level matching then undergo structural matching, which aims to identify potential homonym and synonym conflicts based on the neighbors of the compared concepts. Additionally, structural matching should also recognize hypernym-hyponym pairs and holoynm-meronym pairs. If such conflicts are identified as likely, a respective warning is added to the preliminary matching decisions.

Finally, taxonomy-based matching (e.g., the Lesk metric) can be optionally performed for concept pairs, which are still presumed "independent" after structural matching. The goal is to detect potential hidden relatedness between the concepts. This is especially recommended if at least one of the compared concepts is yet unknown in the domain ontology. The final matching proposals, including any warnings, are presented to domain experts, who then have the chance to accept the proposals or override them. For instance they can decide if and how potential homonymy/synonymy conflicts should be resolved. If no domain expert is available, the default proposals are pursued.

Based on the matching results, specific integration proposals are generated in the schema consolidation phase. In summary, the following strategies are applied [29]: For matching concepts, the integration proposal is to merge them and make sure that both concept names are stored in a repository otherwise this could result in semantic loss [5]. Unrelated concepts are transferred to the integrated schema independently. For (directly) related concepts, both concepts are transferred to the integrated schema and a relationship between them is introduced. Concepts are indirectly related when they have no direct connecting relationship in the domain but are connected via several other concepts. For example two concepts might have a common concept with which they are connected via hypernym-hyponym and holonym-meronym relationships. It is principally possible to also transfer such more complex relationships – including all intermediate concepts – to the integrated schema, as a proper connection for the indirectly related concepts.

A central requirement regarding the integration workflow states that the process should be automatized as much as possible. This means that domain experts should be supported by preferably accurate proposals and the tool should generate a default integrated schema even when no user input is made at all. The integration prototype provides the option to adjust the preferred degree of automatization.

Currently, the prototype focuses on certain matching techniques and was mainly tested for exemplary cases. However the preliminary results give reason to hope that the suggested workflow is a suitable default process for most projects.

V. Summary and Conclusions

In this paper, we have presented a three-tier matching strategy for predesign schema elements. Our strategy is modeling-language independent and should be applied early in the information system development process. Modeling-language independent means that detailed implementation and design information is not dealt with at this stage and that we only use the most essential modeling elements: concepts and links between concepts. Our approach should be viewed as one step towards a semi-automatic method for modeling-language independent schema integration. The presented and proposed multi-level matching strategy is composed of *element level matching*, followed by *structural level matching* and ending with *taxonomy-based matching*. When applied in schema integration, the three matching strategies should facilitate the recognition of similarities and differences between two source schemata.

References

[1] A., Bachmann, W. Hesse, A. Russ, C. Kop, H.C., Mayr, and J. Vöhringer J., "OBSE – An Approach to Ontology-Based Software Engineering in the practice," in EMISA, 2007, pp. 129-142.

[2] S. Banerjee and T. Pederson, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," in Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, 2002, pp. 136–145.

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

[4] P. Bellström, View Integration in Conceptual Database Design – Problems, Approaches and Solutions, Licentiate Thesis, Karlstad University Studies 2006:5, 2006.

[5] P. Bellström, "On the Problem of Semantic Loss in View Integration," in Information Systems Developent Challenges in Practice, Theory, and Education, Vol. 2, C. Barry et al., Eds. Heidelberg: Springer, 2009, pp. 963-974.

[6] P. Bellström, Schema Integration – How to Integrate Static and Dynamic Database Schemata, Dissertation, Karlstad University Studies 2010:13, 2010.

[7] P. Bellström, "A Rule-Based Approach for the Recognition of Similarities and Differences in the Integration of Structural Karlstad Enterprise Modeling Schemata," in Proceedings of the 3rd IFIP WG 8.1 Working Conference on The Practice of Enterprise Modeling, P. van Bommel et al., Eds. Heidelberg: Springer, 2010, pp. 177-189.

[8] Bellström P. and J. Vöhringer, "Towards the Automation of Modeling Language Independent Schema Integration," in Proceedings of the 1st International Conference on Information, Process, and Knowledge Management, A. Kusiak and S.G. Lee, Eds. IEEE Computer Socity, 2009, pp. 110-115.

[9] P. Bellström, J. Vöhringer, and C. Kop, "Towards Modeling Language Independent Integration of Dynamic Schemata," in Information Systems Development Toward a Service Provision Socity, G.A. Papadopoulos et al., Eds. Heidelberg: Springer, 2009, pp. 21-29.

[10] P. Bellström, J. Vöhringer, and A. Salbrechter, "Recognition and Resolution of Linguistic Conflicts: The Core to a Successful View and Schema Integration," in Advances in Information Systems Development New Methods and Practice for the Networked Socity, Vol. 2, G. Magyar et al., Eds. Heidelberg: Springer, 2007, pp. 77-87.

[11] L. Bloomfield, Language, Chicago - London: The University of Chicago Press, 1933.

[12] P. Chen, "The Entity-Relationship Model – Toward a Unified View of Data," ACM Transactions on Database Systems, vol. 1(1), 1976, pp. 9-36.

[13] R. Gustas and & P. Gustiené, "Towards the Enterprise Engineering Approach for Information System Modelling Across Organisational and Technical Boundaries," in Enterprise Information Systems V, O. Camp et al., Eds. Dordrecht: Kluwer, 2004, pp. 204-215.

[14] Q. He and T.W. Ling, "Resolvning Schematic Descrepancy in the Integration of Entity-Relationship Schemas," in: Proceedings of ER 2004, P. Atzeni et al., Eds. Heidelberg: Springer, pp. 245-258.

[15] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in WordNet: An Electronic Lexical Database (Language, Speech, and Communication), 1998.

[16] D. Kensche, C. Quix, X. Li, and Y. Li, "GeRoMeSuite: A System for Holistic Generic Model Mangement," in Proceedings of the 33rd international conference on Very large data bases, C. Kock et al., Eds. 2007, pp. 1322-1325.

[17] M.L. Lee and T.W. Ling, "A Methodology for Structural Conflict Resolution in the Integration of Entity-Relationship Schemas," Knowledge and Information Systems, vol. 5(2), 2003, pp. 225-247.

[18] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," in AAAI'06, 2006, pp. 775-780.

[19] G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, vol. 38, 1995, pp. 39-41.

[20] Object Management Group, OMG Unified Modeling Language (OMG UML), Superstructure, [Electronic], Available: http://www.omg.org/spec/UML/2.2/Superstructure/PDF/ [20101201]

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

[22] T. Pederson, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity - measuring the relatedness of concepts," in Proceedings of the 19th National Conference on Artificial Intelligence, 2004, pp. 1024-1025.

[23] E. Rahm and P.A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," The VLDB Journal, vol. 10, 2001, pp. 334–350.

[24] P. Shvaiko and J. Euzenat, "A Survey of Schema-Based Matching Approaches," Journal of Data Semantics, vol. 4, 2005, pp. 146-171.

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

[26] J. Vöhringer, P. Bellström, D. Gälle, and C. Kop, "Designing a study for evaluating user feedback on predesign models," in Proceedings of ISD2009 Conference, 2010, pp. 411-425.

[27] J. Vöhringer, D. Gälle, G. Fliedl, C. Kop, and M. Bazhenov, "Using Linguistic Knowledge for Fine-tuning Ontologies in the Context of Requirements Engineering," International Journal of Computational Linguistics and Applications, Vol. 1(1-2), 2010, pp. 249-267.

[28] J. Vöhringer and G. Fliedl, "Adapting the Lesk Algorithm for Calculating Term Similarity in the Context of Ontology Engineering," in Proceedings of ISD2010 Conference, 2011, In Print.

[29] J. Vöhringer and H.C., Mayr, "Integration of Schemas on the Pre-Design Level Using the KCPM-Approach," in Advances in Information Systems Development Bridging the Gap between Academia and Industry, A.G. Nilsson et al., Eds. Heidelberg: Springer, 2006, pp. 623-634.

[30] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994, pp. 133-138.

Semantic-enabled Efficient and Scalable Retrieval of Experts

Witold Abramowicz, Elżbieta Bukowska, Monika Kaczmarek, Monika Starzecka
Department of Information Systems, Faculty of Informatics and Electronic Commerce, Poznan University of Economics, Poznań, Poland
{w.abramowicz; e.bukowska; m.kaczmarek; m.starzecka}@kie.ue.poznan.pl

Abstract—**Nowadays, efficient utilization of knowledge became a key to the success of an organization. The need to identify experts within or outside an organization has been for a long time inspiration for various initiatives undertaken by academia and industry. The eXtraSpec system developed in Poland is an example of such initiatives. In order to realize its tasks, the eXtraSpec system needs not only to be able to acquire and extract information from various sources, but also requires an appropriate information representation supporting reasoning over person characteristics. The considered mechanism should allow for precise identification of required data, but simultaneously, be efficient and scalable. The main goal of this paper is to present the reasoning scenario we applied within the eXtraSpec project and discuss the underlying motivation, which led to the development of pre-reasoning mechanism. The system architecture and developed ontology together with implementation details are also discussed.**

Keywords - Expert finding system; knowledge representation; expert characteristic

I. Introduction

Efficient utilization of knowledge is a key to the success of an organization. Knowing the skills and expertise of employees as well as a proper recruitment are of major importance. More and more often organizations take advantage of data available on the Internet to locate experts they require. As the data available is very dispersed and of distributed nature, a need appears to support the human resources management process using IT-based solutions, e.g., information extraction and retrieval systems.

Within an information retrieval (IR) process a single query is executed on a set of resources to identify the relevant data [1]. In general, a typical retrieval system encompasses three components: a module responsible for collecting data and creating its representation in the form of an index; an interface allowing formulating queries consisting of a set of keywords and finally, a mechanism matching a query to the created indexes.

These components affect the quality of the retrieval process i.e., values of the precision and recall metrics.

The traditional expert retrieval systems face well known IR problems caused by application of different keywords and various levels of abstraction by users while formulating queries or by using different words and phrases while creating a description of a phenomenon, based on which indexes are created. Thus, to ensure that in a response to a query an expert retrieval system returns documents, which do not contain words included in the query, but are still relevant, very often semantics is applied.

There are many research and commercial initiatives aiming at development of expert retrieval systems supported by semantics. They are to provide interested parties with detailed information on people's experience and skills. One of such initiatives is the on-going Polish project eXtraSpec [23]. Its main goal is to combine company's internal electronic documents and information sources available on the Internet to provide an effective way of searching experts with competencies in the given field. The system is to support three main scenarios: finding experts with desired characteristic, defining teams of experts and verifying data on a person in question. In order to support these scenarios, the eXtraSpec system needs not only to be able to acquire and extract information from various sources, but also requires an appropriate information representation supporting reasoning over person's characteristics. In addition, the mechanism should allow on the one hand for precise identification of required data and, on the other hand, be efficient and scalable.

The main goal of this paper is to present the reasoning approach we followed within the eXtraSpec project and discuss the underlying motivation, which led to the development of a semantic-based mechanism to retrieve experts in its current state. In addition, the ontology developed to describe expert characteristics is presented.

In order to fulfill the mentioned goals, the paper is structured as follows. First, the related work in the area of expert finding systems is discussed. Next, the ontology developed for the needs of the eXtraSpec project to support retrieval of experts is presented. Then, the short description of the considered scenarios regarding the application of the reasoning infrastructure follows. Finally, the system architecture as well as implementation details are given. The paper concludes with final remarks.

II. Related Work

The need to find expertise within an organization has been for a long time inspiration for initiatives aiming at development of a class of search engines called expert finders [2]. There are several aspects connected with expert finding, for instance, following McDonald and Ackerman

[3] those may be: expertise identification aiming at answering a question: ‘who is an expert on a given topic’ and expertise selection aiming at answering a question ‘what does X know’? Within our research, we focus on the first aspect i.e., identifying a relevant person given a concrete need.

First systems focusing on expertise identification relied on a database like structure containing a description of experts’ skills (e.g., [4]). However, such systems faced many problems, e.g., how to ensure precise results given a generic description of expertise and fine-grained and specific queries [5] or how to guarantee the accuracy and validity of stored information given the static nature of a database. To address these problems other systems were proposed focusing on automated discovery of up-to-date information from specific sources such as e.g., email communication [6]. In addition, instead of focusing only on specific document types, systems that index and mine published intranet documents were proposed [7]. An example may be the Spree project [8] aiming at providing automatic expert finding facility, able to answer a given question. The system automatically builds qualification profiles from documents and uses communities and social software in order to provide efficient searching capabilities. In addition, currently the Web itself offers many other possibilities to find information on experts, as there are a number of contact management portals or social portals where users can look for experts, potential employees or publish their curricula in order to be found by future employers (some examples may be [24] [25], [26]).

When it comes to the algorithms applied to assess whether a given person is suitable to a given task, at first, standard IR techniques to locate an expert on a given topic were applied [9][10]. Usually, expertise of a person was represented in a form of a term vector and a query result was represented as a list of relevant persons. If matching a query to a document relies on a simple mechanism checking whether a document contains the given keywords, then the well-known IR problems occur: (1) low precision of returned results (there is a word, but not in this context); (2) low value of recall (relevant documents described with another set of keywords are not identified); (3) a large number of documents returned by the system (especially in response to a general query) the processing of which is impossible (e.g., due to the time limit). Therefore, few years ago, the Enterprise Track at the Text Retrieval Conference (TREC) was started in order to study the expert-finding topic. It resulted in further advancements of the expert finding techniques and application of numerous methods such as probabilistic techniques or language analysis techniques to improve the quality of finding systems (e.g., [11] [12] [13] [14]).

As the Semantic Web technology is getting more and more popular, it is not surprising that it has been used to enrich descriptions within expert finding systems. The introduction of semantics into search systems may take two forms: the use of semantics in order to analyze indexed documents or queries (query expansion), or operating on semantically described resources (e.g., RDF files) with use of reasoners.

Within the expert finding systems, both approaches have been applied and a number of ontologies to represent competencies and skills was developed. For instance, a Single European Employment Market-Place project [15] aiming at providing interoperable architecture for e-Employment services. used an ontology in order to create a semantic description of job offers and CV. The ontology is called “Reference Ontology” and it consists of thirteen sub-ontologies: Competence, Compensation, Driving License, Economic Activity, Education, Geography, Job Offer, Job Seeker, Labour Regulatory, Language, Occupation, Skill and Time. It was built based on the commonly used standards, e.g., ISO 4217 [27], ISCO-88 (COM[28]), ONET [29], DAML ontology [30].

In turn, in [16], authors describe requirements and a process of ontology creation for the needs of HR management. They developed an ontology used by a meta-search engine searching jobs in job portals [17] and by a university competence management system [18]. The ontology was created in the OWL formalism. It consists of sub-ontologies for competencies, occupations and learning objects. Another example is an ExpertFinder [19] framework enabling application of existing vocabularies in semantically supported systems. It provides terms and best practices for describing web pages, persons, institutions, events, areas of expertise or educational aspects. It uses such vocabularies as: FOAF [31], SIOC [32], vCard [33] or Dublin Core [34].

In addition, numerous ontologies, taxonomies and classifications have been created in the HR management area, e.g., the Standard Occupational Classification (SOC) [35] of the US Federal statistical agencies or taxonomy of skills in the KOWIEN project [20].

The system discussed in this paper relates to the semantic-based expert finding. However, our setup is a bit different. The eXtraSpec system acquires information from outside and assumes that one can build an expert profile based on the gathered information. The system gathers information on a large set of experts. More experts imply bigger topic coverage and increased probability of a question being answered. However, it also causes problems connected with the heterogeneity of information as well as precision and recall of the system. The application of semantics may help to normalize the gathered data and ensure appropriate level of precision and recall; however, it generates problems with scalability and efficiency of the designed mechanisms. In addition, the ontology itself developed for the needs of the eXtraSpec system differs from other projects: (1) it is not limited to hierarchical relations; (2) it has been developed for the Polish language and relate to Polish standards; (3) it has been built in accordance to the Simple Knowledge Organization System (SKOS) [36] standard.

III. Ontologies in the Extraspec Project

One of the most important functionalities of the eXtraSpec system is the identification of persons having the desired expertise. As already discussed, in order to ensure the quality of returned results, the decision to apply Semantic Web technologies to retrieve and describe profiles was taken. The eXtraSpec system acquires automatically data from dedicated sources, both company external and internal ones. The extracted content is saved as an extracted profile (PE), which is an XML file compliant with the defined structure of an expert profile (Figure 2) based on the European Curriculum Vitae Standard [37]. Vocabulary in the extracted content is then processed and normalized using the developed ontology. The result of the normalization process is a normalized profile (PN). Every normalized profile provides information on one person, but one person may be described by a number of normalized profiles (e.g., information on a given person at different points in time or information acquired from different sources). Thus, normalized profiles are analysed and then aggregated, in order to create an aggregated profile (PA) of a person. Finally, the reasoning mechanism is fed with the created aggregated profiles and answers user queries on experts.

The above-mentioned steps impose requirements on the ontology. It should enable semantic annotation of all elements of profiles as well as support the normalization and discovery process. The basic element of the eXtraSpec system is an already mentioned profile of an expert. Each expert is described with series of information, e.g., first and last name, history of education, career history, hobby, skills, obtained certificates. To make the reasoning possible, the following attributes from the profile of an expert should be linked to ontology instances:

- Educational organization – name of organization awarding the particular level of education or educational title;
- Certifying organization – name of an organization that issued particular certificate;
- Client, employer and role – those three attributes are used to describe history of an employment. A single step in the employment history is described as a business relation. Each relation consists of three basic elements: client – employer and a role (i.e., profession) that an expert played in this relation;
- Scope of education – the domain of education (for example: IT, construction, transportation);
- Topic of education – for a higher education description, it will be a name of the specialization, for trainings or courses etc. – their topic;
- Result of education – the obtained title;
- Skill – an ability to do an activity or job well, especially because someone has practiced it;
- Name of a certificate;
- Degree of a skill.

Performed analysis of the requirements imposed on the ontology for the needs of reasoning, concluded with the definition of a set of relations that should be defined:

- *subConceptOf* – to represent hierarchical relations between concepts,
- *isPartOf* – for representation of composition of elements, for example: ability of using MSWord is a part of ability of using MSOffice (however, knowing MSWord does not imply that a person knows the entire MSOffice suit),
- *isRequired* – connection between two concepts, for example: to have a role – doctor, one must have graduated from some medical school.
- *implies* – from one fact, or set of facts, another fact can be concluded. For example, if one has skills A and B, then he also has skill C.

As the result of the conducted analysis of different formalisms and data models, the decision was taken to use the OWL language as the underlying formalisms and the SKOS model as a data model. The criteria that influenced our choice were as follows: (1) relatively easy translation into other formalisms; (2) simplicity of representation; (3) expressiveness of used ontology language; (4) efficiency of the reasoning mechanism. Many knowledge representations, such as thesauri, taxonomies and classifications share some structure elements and are used in similar applications. SKOS gathers most of those similarities and explicitly enables data and technology exchange between different applications. The SKOS data model enables low cost migration that will allow making a connection between existing SKOS and the semantic Internet. Ontologies developed in accordance to the SKOS model can be expressed in any known ontology language. Because of the strong software support and a wide usage of OWL, we decided to use that formalism within our work.

As a result of the conducted work, the data structure was designed having one SKOS ontology with eight concept schemas for each area of interest: Organizations (for organizational organizations, certifying organizations, Employer and Client), SkillName, SkillDegree, Certificate, Role, EducationScope, EducationTopic and EducationResult. While building the ontology for the needs of the eXtraSpec system, a wide range of taxonomies and classifications has been analyzed in order to indentify best practices and effective solutions. As the eXtraSpec system is a solution designed for the Polish market, so is the developed ontology. For instance, during the development of Concept Schemas for Organizations information provided by the Polish Ministry of Science and Higher Education [38] was used while for Role organization the official Polish Classification of Occupations [39] published by the Polish Ministry of Labor and Social Policy was utilised.

IV. Considered Searching Scenarios

One of the most important functionalities of the eXtraSpec system is the identification of persons having the desired expertise. The application of Semantic Web technologies in order to ensure the quality of returned results implies application of a reasoning mechanism to answer user queries. In addition, the strict requirement towards the performance and scalability of the developed system was formulated. Therefore, a design decision needed to be taken on how to apply semantics and at the same time ensure the required quality of the system during the discovery process.

Given the above criteria (precision and recall on the one hand, and efficiency and scalability on the other), three possible scenarios were considered. The *first* scenario involves using the fully-fledged semantics by expressing all expert profiles as instances of an ontology, formulating queries using the defined ontology, and then, executing a query using the reasoning mechanism. This approach involves the need to load all ontologies into the reasoning engine and representing all individual profiles as ontology instances. Performed experiments showed that querying the reasoning infrastructure, even while using only a small set of gathered profiles, is resource (large memory consumption) and time consuming task (up to a few minutes). Therefore, although having a high precision and recall, it has poor performance and scalability.

The *second* scenario relies on query expansion using ontology, i.e., adding keywords to the query by using an ontology to narrow or broaden the meaning of the original query. It allows to get answers faster than the previous scenario, however, it could not take into account additional relations expressed in the ontology, and therefore, did not always allow for increased precision. In addition, each user query needs to be normalized and then expanded using ontology, therefore, application of a reasoner was necessary. The experiments showed that it affected the values of system performance and scalability.

The *third* scenario called pre-reasoning involves two independent processes: creation of enriched profiles (indexes), to which additional information reasoned from the ontology is added and saved within the repository as syntactic data; formulating query with the help of the appropriate GUI using the defined ontology serving as a controlled vocabulary. Then, the query is executed directly on a set of profiles using the traditional mechanisms of IR. There is no need to use the reasoning engine while executing a query. This approach allows circumventing the drawbacks associated with the first approach, shifting the burden of an operation on the stage of indexing using ontologies.

Summarizing, our experiments proved that applying fully-fledged semantics is a precise but neither efficient nor scalable solution. Query expansion provides increased precision of the results (in comparison to traditional IR mechanisms) and has better scalability and efficiency than the fully-fledged semantics, however, does not allow to take full advantage of the developed ontologies and existing relations between concepts. Only application of the third considered scenario allows taking advantage of the mature IR mechanisms while increasing the accuracy and completeness of the returned results by: introducing a preliminary stage called pre-reasoning in order to create enriched indexes and the minimum use of the reasoning engine during the search.

The short overview of the constituents of the proposed mechanism follows.

V. Semantic-enabled Retrieval of Experts

The eXtraSpec system consists of a number of modules specialized for different tasks. Its architecture is described in [22], here we focus on the REA component (REAsoning) presented in Figure 1. It consists of indexing mechanism (indexer), searching mechanism (searcher), composition mechanism (composer) and a reasoning engine with set of ontologies loaded.

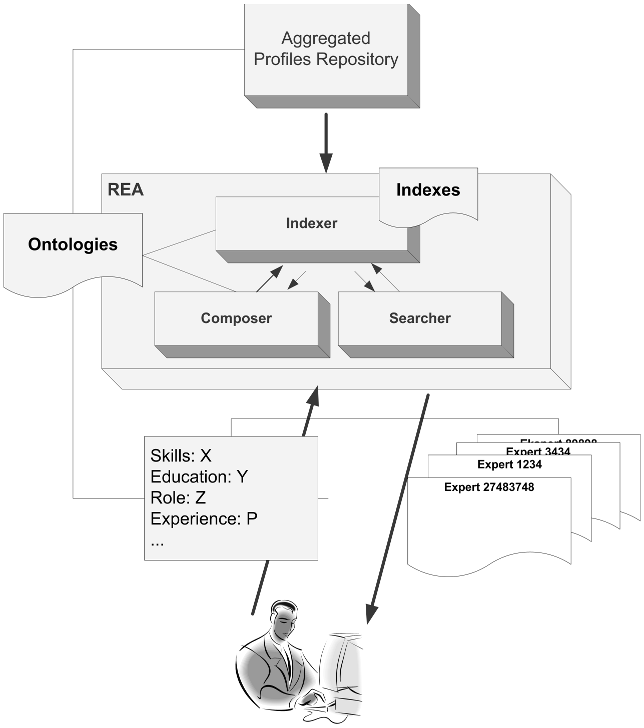


Figure 1. REA overview

The selected scenario requires the support of two independent processes. First, creating profiles' indexes optimized for search, i.e., structured so as to enable a fast search based on criteria preset by a user, and enriched with additional information using an ontology (pre-reasoning).

The second process that needs to be supported is defining the query matching mechanism on the enriched indexes - this process is initiated by a user formulating queries using a graphical interface.

To perform the IR side of the mechanism, the open-source java library Lucene [40], supported by the Apache Software Foundation, was selected. Fields in the Lucene documents cannot be grouped together nor stored as hierarchical structures. However, within an aggregated profile (PA), which is a base profile for searching, some hierarchies and groups might be found. Since explicit mapping from PA to the Lucene document is not possible, during the indexing process profiles are divided into a number of separate documents as shown in Figure 2. Concurrently with the indexing process, pre-reasoning takes place, in order to complete profile with implied facts. Documents contain fields generated directly from PA (marked with +) as well as the additional fields (marked with #). In the simplest case, the reasoning engine returns a list of all superconcepts for the given element to the indexing module. The hierarchy of superconcepts is preserved. Superconcepts are indexed as additional values for the given document field: these values are saved as next array elements and it is assumed that the higher array index number, the smaller weight the concept has. The assigned weight affects the ranking procedure. If returned superconcepts do not correspond with PA elements conceptually, additional fields are added to the document being indexed. For example, PA element „address" might be divided into data that are more detailed, i.e., zip code, city, street, etc. Based on the zip code it is possible to specify county and province and search for experts using the spatial criteria. Since PA does not contain such elements, we add fields to the personal data document during the indexing process. In turn, documents that contain information about education history have been expanded by field „catOfEduOrganization". Hierarchy of superconcepts for each education organization is acquired from ontology and indexed. This enables, e.g., searching people who graduated from a desired type of educational organization, e.g., any technical university. In the document that contains data about skills, the field „skillName" was expanded in order to contain all superconcepts from the ontology for a given skill. The same expansion was made in the document about history of employment, were the field „role" was expanded using the hierarchy of superconcepts for occupied positions.

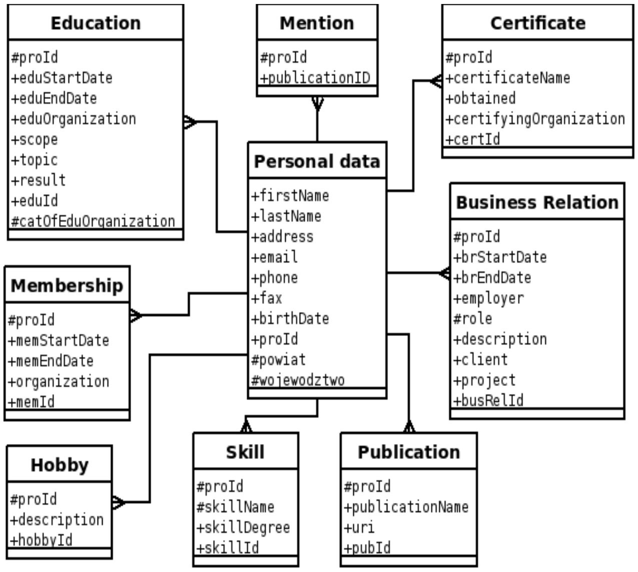


Figure 2. Data model overview

Lucene provides a very flexible but simple query structure. Therefore, in the eXtraSpec system it had to be extended in order to correspond to the PA structure and searching scenarios. In order to execute more sophisticated queries encompassing several criteria from various documents, a QueryObject structure needed to be defined. It stores information on fields' names, required values as well as logical operator that should be fulfilled.

The performed tests have shown that the system fulfils the defined requirements. Application of semantics in the form of a pre-reasoning phase allowed to achieve precise results, simultaneously allowing to take advantage of the matured IR mechanisms guaranteeing scalability and good performance of the system.

VI. Conclusions and Future Works

The main goal of the eXtraSpec project is to develop a system supporting analysis of company documents and selected Internet sources for the needs of searching for experts from a given field or with specific competencies. The provided system focuses on processing texts written in the Polish language. The obtained information is stored in the system in the form of experts' profiles and may be consolidated when needed. The system aims to offer a user friendly interface to perform queries that allow to find persons with specific characteristics. Realisation of this goal requires interconnection between developed interface and underlying ontologies. Within this paper, we have discussed the concept and considered scenarios regarding the implementation of the reasoning mechanism for the needs of the eXtraSpec system. We argue that by introducing the pre-reasoning phase, the application of semantics may be used to achieve precise results when searching for experts and at the same time, ensure the proper performance and scalability.

The set of developed ontologies discussed within this paper was designed specially for the Polish language, however, the main structure and model as well as defined relations may be reused also for other languages. As mentioned, the ontology in question is still under development, however, in the current state of affairs the reasoning about competencies in order to complete person's profile with additional data on education, work experience is successfully performed by the REA component described within this paper. Our future work focuses on the implementation of the second scenario supported by the eXtraSpec system i.e., composition of teams of experts using the developed ontology.

Acknowledgment

The work published in this article was supported by the project titled: "Advanced data extraction methods for the needs of expert search"financed under the Innovative Economy Framework and partially supported by European Union in the European Regional Development Fund programme (contract no. UDA-POIG.01.03.01-30-150/08-01).

References

[1] van Rijsbergen, C. J.; “Information Retrieval and Information Reasoning”. Computer Science Today 1995,pages 549-559

[2] Yimam, D.; “Expert finding systems for organizations: Domain analysis and the demoir approach” in: ECSCW 999 Workshop: Beyond KNowledge Management: Managing Expertise, pages 276–283, New York, NY, USA, 1996. ACM Press

[3] McDonald, D. W. and Ackerman, M. S.; “Expertise recommender: a flexible recommendation system and architecture” in: CSCW ’00: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 231–240. ACM Press, 2000.

[4] Yimam-Seid, D. and Kobsa, A. “Expert finding systems for organizations: Problem and domain analysis and the demoir approach”. Journal of Organizational Computing and Electronic Commerce, 13(1):1–24, 2003

[5] Kautz, H., Selman, B., and Milewski, A.; “Agent amplified communication” in: Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pages 3–9, 1996

[6] Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B.; “Expertise identification using email communications” in: CIKM ’03: Proceedings of the twelfth international conference on Information and knowledge management, pages 528–531. ACM Press, 2003

[7] Hawking, D.; “Challenges in enterprise search” in: Proceedings Fifteenth Australasian Database Conference, 2004

[8] Metze, F., Bauckhage, Ch., and Alpcan, T., “The "Spree" Expert Finding System“ in: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA, pp. 551--558

[9] Ackerman, M.S., Wulf, V. and Pipek, V.; “Sharing Expertise: Beyond Knowledge Man-agement”; MIT press, (2002).

[10] Krulwich, B. and Burkey, C.; “ContactFinder agent: answering bulletin board questions with referrals” in: Proceedings of the National Conference on Artificial Intelligence, pages 10-15, 1996

[11] Balog, K., Azzopardi L. and De. Rijke, M.; “Formal models for expert finding in enterprise corpora” in: Proceedings of the ACM SIGIR, pages. 43-50, 2006.

[12] Fang, H. and Zhai, C.; “Probabilistic models for expert finding” in: Proceedingsof the ECIR, pages 418-430, 2007

[13] Petkova, D. and Croft, W.; “Hierarchical language models for expert finding in enterprise corpora” in: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, pages 599-608, 2006

[14] Serdyukov, P. and Hiemstra, D.; “Modeling documents as mixtures of persons for expert finding” in: Proceedings of the ECIR, pages 309-320, 2008.

[15] Gómez-Pérez, A., Ramírez, J., and Villazón-Terrazas, B., “An Ontology for Modelling Human Resources Management Based on Standards” in: B. Apolloni et al. (Eds.): KES 2007/WIRN 2007, Part I, LNAI 4692, pp. 534–541, 2007

[16] Dorn, J., Naz, T., and Pichlmair, M., “Ontology Development for Human Resource Management” in: "Proceedings of 4rd International Conference on Knowledge Management", Ch. Stary, F. Barachini, and S. Hawamdeh (Hrg.); Series on Information&Knowledge Management, 6 (2007), ISBN: 978-981-277-058-5; pp. 109 - 120.

[17] Dorn, J. and Naz, T.; “Meta-search in Human Resource Development”, in: Proceedings of 4th Int. Conference on Knowledge Systems, Bangkok, Thailand, 2007

[18] Dorn, J. and Pichlmair, M.; “A Competence Management System for Universities”, in: European Conference on Information Systems, St. Gallen, 2007, pp.759 - 770

[19] Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J.G., Mochol, M., Nixon, L.JB., Polleres, A., and Zhdanova, A.V., “Combining RDF Vocabularies for Expert Finding”. In: Proceedings of the 4th European conference on The Semantic Web: Research and Applications ESWC '07, 2007, pp. 235--250

[20] Dittmann, L.;“Towards Ontology-based Skill Management, Projektbericht zum Verbundprojekt KOWIEN”, Universität Duisburg-Essen, 2003.

[21] Abramowicz, W., Wieloch, K.; “Raport podsumowujący wyniki prac przeprowadzonych w ramach zadań Z1.1, Z1.2 oraz Z2.1”, Technical report of the eXtraSpec project, Department of Information Systems, Poznan University of Economics, 2009

[22] Abramowicz, W., Kaczmarek, T., Stolarski, P., Węcel, K., and Wieloch, K.; “Architektura systemu wyszukiwania ekspertów eXtraSpec”, in: Proceedings of “Technologie Wiedzy w Zarządzaniu Publicznym”, Hucisko, 19-21 September 2010

[23] http://extraspec.kie.ue.poznan.pl/, last access date: 3.12.2010

[24] http://www.bizwiz.com, last access date:3.12.2010

[25] http://www.xing.com, last access date: 3.12.2010

[26] http://linkedin.com. last access date: 3.12.2010

[27] http://www.iso.org/iso/en/prods-services/popstds/currencycodeslist.html, last access date: 3.12.2010

[28] http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC, last access date: 3.12.2010

[29] http://online.onetcenter.org/, last access date: 3.12.2010

[30] http://cs.yale.edu/homes/dvm/daml/time-page.html, last access date: 3.12.2010

[31] http://www.foaf-project.org/, last access date: 3.12.2010

[32] http://sioc-project.org/, last access date: 3.12.2010

[33] http://www.imc.org/pdi/, last access date: 3.12.2010

[34] http://dublincore.org/, last access date: 3.12.2010

[35] http://www.bls.gov/soc, last access date: 3.12.2010

[36] http://www.w3.org/TR/swbp-skos-core-spec, last access date: 3.12.2010

[37] http://www.europa-pages.com/jobs/europass.html, last access date: 3.12.2010

[38] http://www.nauka.gov.pl/szkolnictwo-wyzsze/system-szkolnictwa-wyzszego/uczelnie/, last access date: 3.12.2010

[39] http://www.praca.gov.pl/pages/klasyfikacja_zawodow2.php, last access date: 3.12.2010

[40] http://lucene.apache.org, last access date: 3.12.2010

Enhancing Knowledge Flow by Mediated Mapping Between Conceptual Structures

Peteris Rudzajs
Institute of Applied Computer Systems
Riga Technical University
Riga, Latvia
peteris.rudzajs@rtu.lv

Marite Kirkova
Department of System Theory and Design
Riga Technical University
Riga, Latvia
marite.kirikova@cs.rtu.lv

***Abstract*. Intra- and inter-institutional knowledge flow usually is hindered by a number of mutually related knowledge barriers. Removing some of these barriers may enhance knowledge flow and thus open new opportunities for cooperation. In this paper we illustrate how indirect conceptual mappings supported by a software tool can remove some of knowledge flow barriers and have a positive impact on the knowledge flow. The approach described in the paper is presented in the context of cooperation between industrial and educational institutions in the area of information and communication technologies.**

Keywords- knowledge flow, concept structures, competence framework; mediated comparison; study courses; job positions

I. INTRODUCTION

Knowledge flow is a process whereby knowledge is passed between people or knowledge processing mechanisms [1], [2]. There are a number of barriers that may hinder the flow of knowledge. In [2] these barriers have been grouped into 5 categories: knowledge characteristics (such as causal ambiguity and non-valid knowledge), knowledge source barriers, knowledge receiver barriers, contextual barriers, and insufficient mechanisms. This paper addresses only two of these categories, namely, knowledge receiver barriers and contextual barriers. The knowledge receiver barrier is minimized by raising absorptive capacity of knowledge flow receivers. Insufficient mechanism is addressed by proposing an information technology solution that facilitates knowledge exchange between two parties. The issues are discussed in the context of Information and Communication Technology (ICT) knowledge exchange between industrial organizations and educational institutions operating in European Union (EU) countries. The main focus is on the fit between knowledge/competence/skills demand of industrial organizations and knowledge/competence/skills offer by educational institutions.

Industrial organizations (further in the text - Industry) usually maintain the so-called job position framework that consists of a list of positions and their descriptions in terms of responsibilities and competencies. On the other hand educational institutions (further – Universities) maintain course descriptions in terms of topics covered, learning outcomes, obtainable knowledge, and skills. Different concept systems are used for the description of Industry job positions and University courses. This causes a knowledge receiver barrier - low absorptive capacity (lack of sufficient related knowledge to assess the value of transferred knowledge) for both parties in exchanging competence knowledge. Because of this barrier University and Industry find it difficult to “understand” each other [3]. Taking into consideration that competence is becoming a kind of “currency” in the job market [4], there is a need to compare competence demand from Industry and competence offer from University in order to see if the university can satisfy competence needs of Industry. This requires improving absorptive capacity of both partners. The purpose of this paper is to show that comparison of competence demand and offer becomes possible (the absorptive capacity barrier can be lowered or removed) if standardized competence frameworks are utilized as a mediating conceptual structure between “languages” of Industry and University. The use of the framework is practically possible only if a supporting information technology is available. In other words – addressing the absorptive capacity as a cause of receiver barrier requires addressing the insufficient mechanism barrier, too.

The method proposed in this paper extends the mediated comparison method described in [5] by utilization of several mediating frameworks. The extension is made in order to facilitate knowledge exchange not only about competencies, but also about ICT tools and technologies used in University and Industry. In the paper the main emphasis is on a developed prototype supporting mediated competence comparison with respect to Industry job descriptions and University courses. The prototype is a part of a collaboration support system that has been designed with the purpose to maintain and exchange information between University and Industry [3], [6], [7], [8], [9]. It helps to remove another Industry-University knowledge flow barrier, namely, the lack of mechanism for knowledge exchange [7].

The paper is structured as follows. Related work is briefly discussed in Section II. In Section III, we describe how conceptual structures which have a potential to be absorbed by actors (people and technologies) of knowledge flow [2] were identified. In Section IV, the method of mediated mapping between conceptual structures is discussed. In Section V, the prototype that is used for enhancing knowledge flow is described. In Section VI, preliminary results are presented. In Section VII, expected contributions, research limitations and some directions of future work are presented.

II. Related work

Research in the field of knowledge exchange between University and Industry has already been done by developing the architecture of University-Industry collaboration support system and its services [3], [6], [7], [8], [9]. The architecture identifies the main areas of action to support knowledge exchange, i.e., (1) knowledge acquisition services, (2) study course services (including services for developing standardized study course descriptions), (3) knowledge representation services, (4) repository services, and (5) analysis services for the analysis of collected information. The architecture incorporates the following approaches to bridge the gap between University and Industry [8]: (1) the use of standardized competence frameworks, (2) the use of tools that automatically interpret ("translate") competences expressed in Industry terms into competences expressed in University terms, (3) giving an opportunity for industry to evaluate technology-oriented elective courses directly or indirectly (course evaluation prototype has been developed), (4) equipping Industry with University insights in skill development trends. The "gap" here means a difference between University and Industry in understanding the essence of knowledge, skills, and competences. The evaluation of technology-oriented elective courses revealed that Industry finds it hard to understand the terms used in course descriptions, thus in this paper we focus on knowledge exchange about knowledge/competence/skills demand and knowledge/competence/skills offer by utilizing (2), (3) and (5) areas of action: (2) by means of study course description with standardized competencies, (3) by using standardized competence frameworks and (5) by using prototype to identify courses corresponding to specific job position (see Section V).

Since the knowledge representation services is the core for maintaining knowledge flow between University and Industry in the architecture of the collaboration support system, some standards for representing knowledge should be selected. In the case of competence information exchange, competence standards (frameworks) should be considered. Several competence frameworks have been developed by different professional and academic organizations and societies, such as European e-Competence framework (e-CF, developed in the European Union) [10], Skills Framework for Information Age (SFIA, developed in the United Kingdom) [11], Club Informatique des Grandes Entreprises Françaises framework for job profiles (CIGREF, developed in France, a short description available in [12]), Advanced IT Training System (AITTS, developed in Germany, a short description available in [12]) as well as curriculum models developed by ACM [13]. Some of these frameworks can be mapped both to the descriptions of University courses and job positions. Based on analysis of competence frameworks [5], [6] the e-CF was selected as the most appropriate framework because e-CF has from cooperation between representatives of Industry and University of several EU countries, therefore it can relatively easily be absorbed by Industry and University operating inside boundaries of EU [5], [14], [15]. Studies of related work did not reveal any approaches that would try to obtain mappings between University courses and Industry job positions by incorporating some standardized competence frameworks. In this paper we explain how mediated mapping is used to map courses and job positions.

Domain ontologies are commonly used for representing conceptualizations [16]. If the job positions and study courses are represented as ontologies and we intend to identify how they are related, then mapping between ontologies should be established. Before establishing the mapping we propose to add the competence context [17] that states that mapping is done based on competencies, tools, and technologies required for job position and acquired in University courses. As a result, mapping between job positions and study courses is becoming indirect. Instead of using ontologies (the approach still requires deeper research of ontology matching problems [18]), we propose to use hierarchical conceptual structures representing controlled vocabularies [19] for job position frameworks, university study programs, tools and technologies, and competence frameworks. This leads to the use of simpler hierarchical structures which are easier to compare by means of initial mapping between element values of conceptual structures (in further text we use "mapping of conceptual structures"). Because Industry does not maintain their own ontologies to formally define knowledge and skills, standardized competence frameworks were used to facilitate initial knowledge exchange and to obtain initial mapping between courses and job positions.

III. Identification of conceptual structures

Basic conceptual structures relevant to University and Industry include: (1) a job position framework (which provides systematization of job descriptions), (2) University study programs and (3) a competence framework (e-CF in this case). Describing a job position in Industry, information about knowledge of existing tools and technologies (T&T) is often included in descriptions. A similar situation is in University – knowledge about specific T&T is included in course content. Therefore it is relevant to include information about T&T in course descriptions. Due to the fact that competence frameworks do not include information about specific T&T used in particular competence, the development of T&T catalogue or selection of existing one should be considered in order to bring conceptually closer the descriptions of job positions and University courses. Several catalogues for describing T&T are available, e.g., Google directory, Yahoo directory, O*net Resource Centre [20] tools & technologies etc. We have selected a catalogue provided by the O*net Resource Centre for its simplicity that is a very important feature of a catalogue used for developing basic approach for mapping conceptual structures. It should

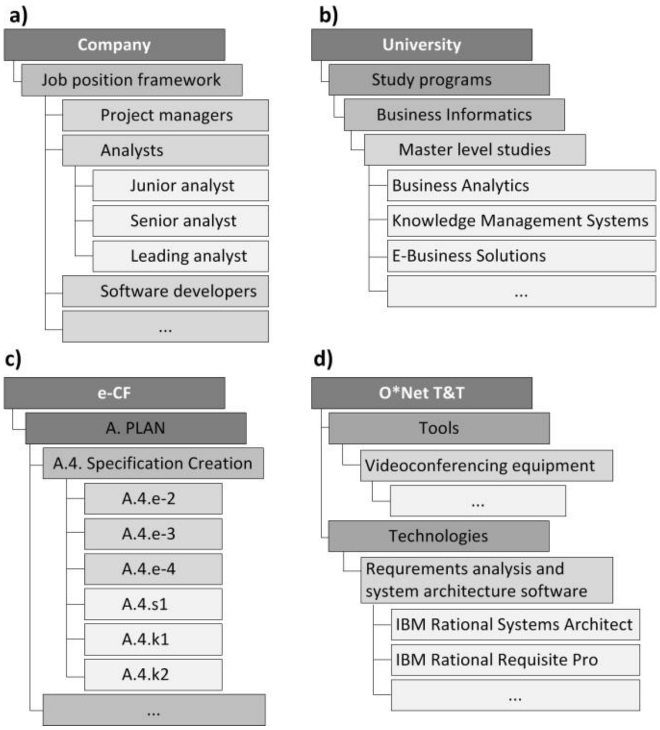


Figure 1. Examples of hierarchical conceptual structures of a) job position framework, b) study program, c) comptence framework and d) T&T catalogue

be mentioned that O*net has developed a list of job positions and the possible corresponding T&T used in the position. We have filtered out of it T&T used in job positions of ICT domain.

The basic conceptual structures mentioned at the beginning of this Section (a total of 4) usually are hierarchical (see Fig. 1) and this structural similarity is utilized in mapping study courses and job positions to corresponding competence framework and T&T catalogue.

IV. Mediated mapping between conceptual structures

After the main conceptual structures (see Fig. 1) are identified, the next step is to consider how these structures could be mapped inside the organization to describe University study courses and Industry job positions. In general, two sets of structures are proposed. The first set (SET1) consists of conceptual structures used in Industry, and the second set (SET2) is used in University. Various competence frameworks and T&T catalogues can be used in organization, but in this paper we assume that Industry and University use the same conceptual structures for the description of job positions and courses (namely, e-CF and O*net T&T).

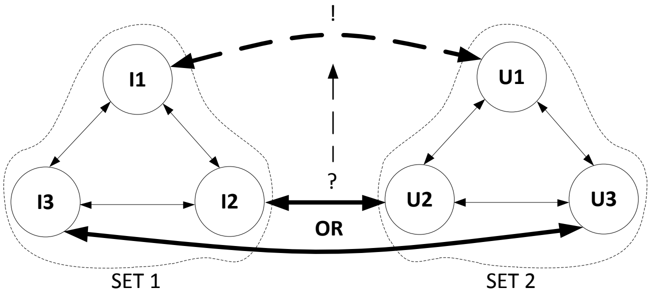


Figure 2. Illustration of basic approach. I – conceptual structures used in Industry, U – conceptual structures used in Univesity

TABLE I. Mapping options

Nr.	Option	Explanation
1	Direct mapping between job positions and university courses (see I1←→U1 in Fig. 2)	This option of direct mapping is not considered as useful because we intend to use mediating hierarchical structures to obtain mapping between job positions and study courses indirectly ensuring the actuality of indirect mapping.
2	Direct mapping between competence frameworks used in University and Industry (see I2←→U2 in Fig. 2)	Competence frameworks are used as mediating structures in order to indirectly map job descriptions to study courses.
3	Direct mapping between T&T catalogues used in University and Industry (see I3←→U3 in Fig. 2)	T&T catalogues are used as mediating structures in order to indirectly map job descriptions to study courses.
4	Direct mapping of both competence frameworks and T&T catalogues used in University and Industry (see I2←→U2 and I3←→U3 in Fig. 2)	Competence frameworks and T&T catalogues are used as mediating structures in order to indirectly map job descriptions to study courses.

Considering that SET1 consists of such elements as a job position framework, e-CF and T&T catalogue, the following mappings between the elements can be introduced (see SET 1 in Fig. 2):

- Job position framework is mapped to e-CF (I1←→I2 in Fig. 2) because every position in the organization is described in terms of standardized competences;
- Job position framework is mapped to T&T (I1←→I3 in Fig. 2) because every position in the organization requires the knowledge of some tools and technologies;
- e-CF is mapped to T&T (I2←→I3 in Fig. 2) because potentially every competence requires some knowledge of tools & technologies.

The same options are in SET2.

Mapping between elements of SET1 and SET2 is necessary in order to identify courses relevant for a particular job position (and v.v.). The possible mapping options and explanation are considered in Table I. Options 2 - 4 identify the possible mediating conceptual structures that are needed to detect the courses relevant for a particular job position. “Mediating conceptual structure” implies that indirect mapping between a job position and university course is based on other conceptual structures, such as competence frameworks and T&T catalogues. Taking into consideration that different organizations can prefer different conceptual structures to formalize knowledge and skills, four mapping options have been identified to demonstrate that fact. The list of options can be extended by the needs of a particular organization. Further in the paper the focus is on option 2 because of emerging value of competences required by Industry and offered by University. Options 3 and 4 are under investigation and are not discussed in this paper in detail.

V. THE PROTOTYPE FOR MEDIATED MAPPING

The method discussed in Section IV is tested by implementing a prototype that serves as one of mechanisms for knowledge exchange between University and Industry.

The prototype has the following basic functionality:

- **Management of organizations –** prototype allows managing information about various organizations that attempt to collaborate in the context of competence information exchange (for example, Universities, partners from Industry).
- **Management of users** - prototype allows managing prototype users belonging to available organizations.
- **Management of hierarchical conceptual structures –** prototype allows managing hierarchical structures such as job position frameworks, University study programs, competence frameworks.
- **Management of mapping -** prototype allows to define mapping between all hierarchical structures (example of mapping study courses to e-CF illustrated in Fig. 4).
- **Establishment of mediated mapping –** mediated mapping between information sources such as job positions and study courses based on mediating conceptual structure e-CF (see Fig. 5).

In order to indirectly map job positions and University courses the following mapping of conceptual structures should be presented:

- Mapping of job positions to Industry competence framework (Nr. 1 in Fig. 3).
- Mapping of study courses to University competence framework (Nr. 2 in Fig. 3).
- Mapping of the competence frameworks of Industry and of University (Nr. 3 in Fig. 3).

When the mapped structures (Industry job positions to a

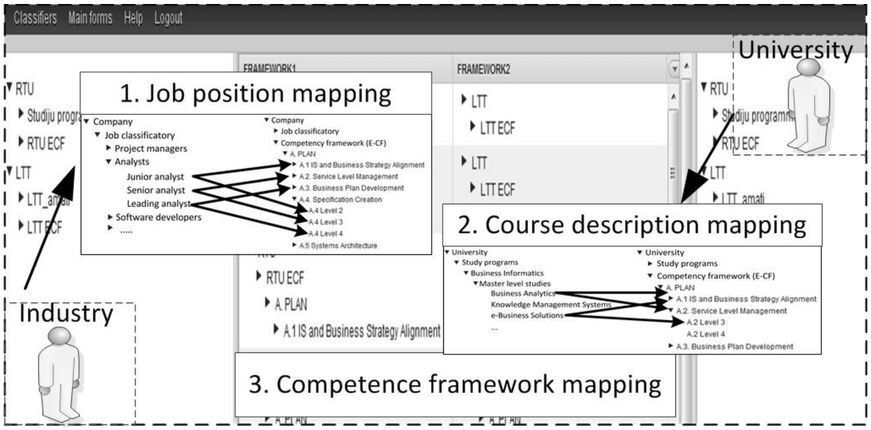
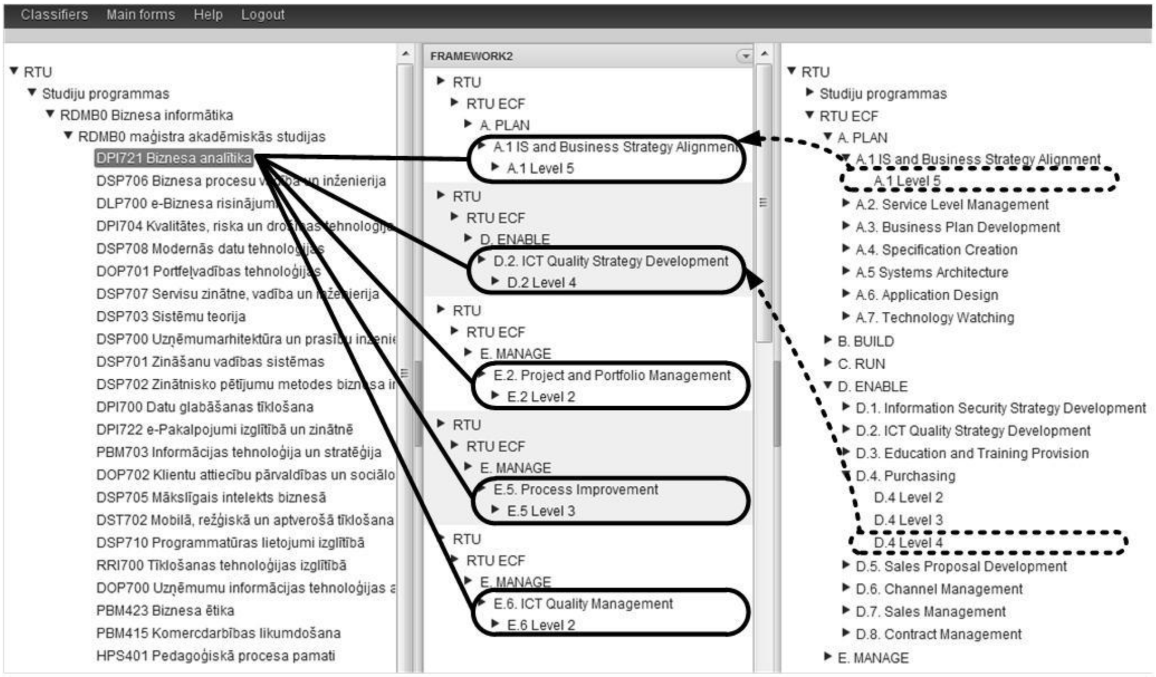


Figure 3. Job position and university course mapping to competence frameworks using a common tool



▼ RTU
▼ Studiju programmas
▼ RDMB0 Biznesa informātika
▼ RDMB0 maģistra akadēmiskās studijas
DPI721 Biznesa analītika
DSP706 Biznesa procesu vadība un inženierija
DLP700 e-Biznesa risinājumi
DPI704 Kvalitātes, riska un drošības tehnoloģija
DSP708 Modernās datu tehnoloģijas
DOP701 Portfeļvadības tehnoloģijas
DSP707 Servisu zinātne, vadība un inženierija
DSP703 Sistēmu teorija
DSP700 Uzņēmumarhitektūra un prasību inženie
DSP701 Zināšanu vadības sistēmas
DSP702 Zinātnisko pētījumu metodes biznesa ir
DPI700 Datu glabāšanas tīklošana
DPI722 e-Pakalpojumi izglītībā un zinātnē
PBM703 Informācijas tehnoloģija un stratēģija
DOP702 Klientu attiecību pārvaldības un sociālo
DSP705 Mākslīgais intelekts biznesā
DST702 Mobilā, režģiskā un aptverošā tīklošana
DSP710 Programmatūras lietojumi izglītībā
RRI700 Tīklošanas tehnoloģijas izglītībā
DOP700 Uzņēmumu informācijas tehnoloģijas a
PBM423 Biznesa ētika
PBM415 Komercdarbības likumdošana
HPS401 Pedagoģiskā procesa pamati

FRAMEWORK2
▶ RTU
▶ RTU ECF
▶ A. PLAN
▶ A.1 IS and Business Strategy Alignment
▶ A.1 Level 5
▶ RTU
▶ RTU ECF
▶ D. ENABLE
▶ D.2. ICT Quality Strategy Development
▶ D.2 Level 4
▶ RTU
▶ RTU ECF
▶ E. MANAGE
▶ E.2. Project and Portfolio Management
▶ E.2 Level 2
▶ RTU
▶ RTU ECF
▶ E. MANAGE
▶ E.5. Process Improvement
▶ E.5 Level 3
▶ RTU
▶ RTU ECF
▶ E. MANAGE
▶ E.6. ICT Quality Management
▶ E.6 Level 2

▼ RTU
▶ Studiju programmas
▼ RTU ECF
▼ A. PLAN
▼ A.1 IS and Business Strategy Alignment
A.1 Level 5
▶ A.2. Service Level Management
▶ A.3. Business Plan Development
▶ A.4. Specification Creation
▶ A.5 Systems Architecture
▶ A.6. Application Design
▶ A.7. Technology Watching
▶ B. BUILD
▶ C. RUN
▼ D. ENABLE
▶ D.1. Information Security Strategy Development
▶ D.2. ICT Quality Strategy Development
▶ D.3. Education and Training Provision
▼ D.4. Purchasing
D.4 Level 2
D.4 Level 3
D.4 Level 4
▶ D.5. Sales Proposal Development
▶ D.6. Channel Management
▶ D.7. Sales Management
▶ D.8. Contract Management
▶ E. MANAGE

Figure 4. Mapping study courses to e-CF

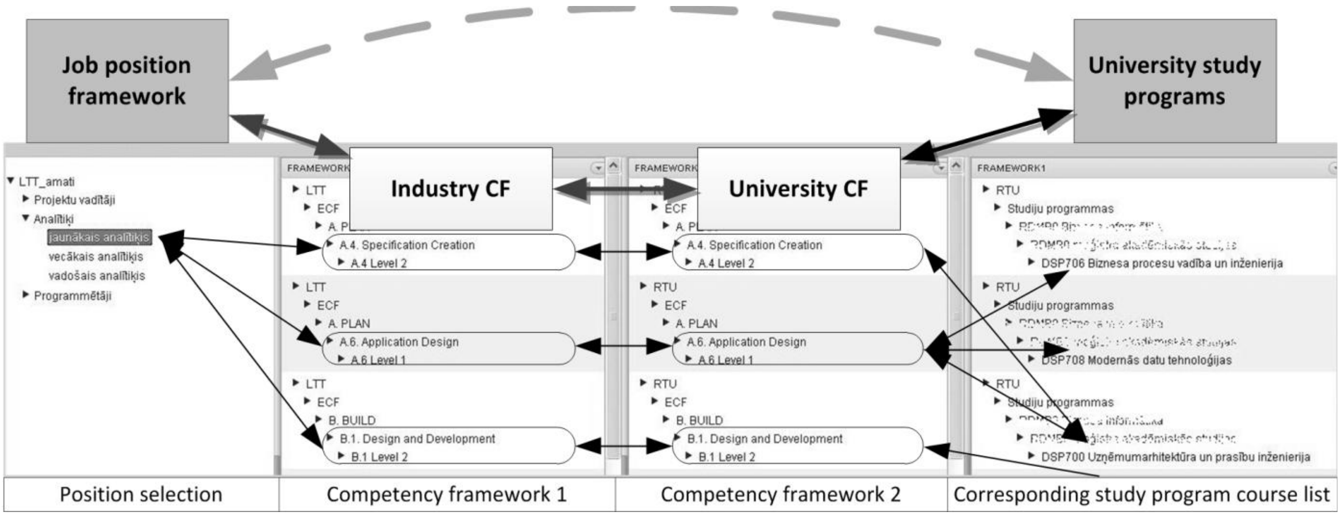


Figure 5. Establishment of mediated mapping

competence framework and University study courses to a competence framework) are available, the next step is to establish mapping between competence frameworks. In this paper we illustrate the case when both University and Industry are using e-CF as a competence framework therefore mapping 1:1 was done using the developed prototype of the tool (see Fig. 3) available for Industry and University.

We consider an example of knowledge exchange about courses corresponding to a particular job position based on e-CF as the mediating structure. Competences required in the job position of Industry job position framework "junior analyst" are mapped to the following e-CF competencies (see Fig. 5): *Specification creation Level 2, Application design Level, Design and development Level 2, Testing Level 1, and Process improvement Level 3.*

The list of courses that include at least one of the requested competences is as follows: Business Analytics; Business Process Management and Engineering; Advanced Data Technologies; Service Science, Management, and Engineering; Enterprise Architecture and Requirements Engineering; Customer Relationship Management and Social Network Technologies; and Artificial Intelligence in Business (in Fig. 5 the course titles are in Latvian).

VI. PRELIMINARY RESULTS

Several representatives of Industry and University were asked to evaluate the prototype. In order to prepare the prototype for evaluation, representatives of one Industry were instructed how to use the prototype to produce mapping between competence framework and job positions/courses in Industry. The same was done on the university side. Afterwards the competence frameworks used in University and Industry were mapped (in the current situation when the representative from Industry and University uses the same competence framework, mapping between these frameworks is 1:1; the developed method and prototype can also support the case when University and Industry use different mediating competency frameworks). This mapping was used to demonstrate the potential of the prototype to show linkage between particular job positions and corresponding study courses.

The results of evaluation revealed the following impacts of the prototype on knowledge exchange between University and Industry:

- Use of a standard mediating conceptual structure improves absorptive capacity of both partners, University as well as Industry.
- Use of IT support brings in transparency in knowledge exchange and considerably shortens the time of comparing the knowledge/competence/skills demand and offer.
- Removal (at least partial) of the above-mentioned knowledge flow barriers: (1) lack of absorptive capacity and (2) insufficient mechanism lower several other knowledge barriers [2] such as causal ambiguity, non-validated knowledge, lack of motivation, unawareness at both ends of knowledge flow, etc.

These results show that the use of the prototype may enhance cooperation between University and Industry if it is systematically used for tuning study programs and developing demand-based courses according to industrial needs.

VII. EXPECTED CONTRIBUTIONS, RESEARCH LIMITATIONS AND FUTURE RESEARCH AVENUES

The method and the prototype presented in this paper were developed to help to remove some barriers of knowledge flow in University and Industry knowledge exchange about knowledge/competence/skills demand and knowledge/competence/skills offer. It was expected that (1) a mutually understandable, internationally recognized competence framework as the mediating conceptual structure will improve absorptive capacity of knowledge to be exchanged and (2) information technology support will make knowledge exchange less time consuming and more transparent.

Preliminary results obtained from prototype evaluation show that ICT supported mediated mapping between conceptual structures can lower or even remove several knowledge flow barriers.

This research is limited to two knowledge flow barriers only. Hypothetically there is a possibility to improve the method and the prototype to address other knowledge barriers. Additional research in knowledge flow barriers is needed to do this. The discussion in this paper is based on the classification of knowledge flow barriers used in domain of healthcare [2]. Experiments with the prototype revealed that the spectrum of and dependencies between knowledge barriers in ICT sector might differ from those in healthcare. Therefore, in order to target properly further investigations, it is necessary to analyze deeper cause consequence relationships between knowledge barriers in the specific area of application – domain of knowledge exchange between University and Industry in ICT sector. Another direction of further research is automatic mapping based on string similarity [21] incorporated in mapping of hierarchical structures.

ACKNOWLEDGMENT

The method and prototype were developed with the support of Lattelecom Technology Ltd. and Riga Technical University in the context of research project N^{o} ZP-2009/15 “Development of the method and the prototype for the normalization and linkage of computer-based competence descriptions”.

REFERENCES

[1] Ch. Lin and J-Ch Wu, “Exploring the influencing factors on inertia source of knowledge flow” in the Proceedings of the 11th International Conference on Electronic Commerce, P. Y. K. Chau, K. Lyytinen, Ch-P. Wei, Ch.C. Yang, and F-R. Lin (Eds), 2009, Tapei, Thailand, pp. 249-258.

[2] C. Lin, B. Tan, and S. Chang, "An exploratory model of knowledge flow barriers within healthcare organizations," Information & Management, vol. 45, no. 5, pp. 331-339, July2008.

[3] P. Rudzajs, L. Penicina, M. Kirikova, and R. Strazdina, "Towards Narrowing a Conceptual Gap Between IT Industry and University," Scientific journal of Riga Technical university, Applied computer systems, vol. Computer Science. Applied Computer Systems - 2010, pp. 9-16, 2010.

[4] M. Jarrar, D. Maynard, J. Hoppenbrouwers, and H. Zhiisheng, "Ontology Outreach to Industry," Knowledge Web Consortium, 2007, pp. 1-81.

[5] P. Rudzajs and M. Kirikova, "Mediated competency comparison between job descriptions and university courses," Scientific journal of Riga Technical university, Applied computer systems, 2011.(in press)

[6] P. Rudzajs, "Development of educational institution and employer collaboration support system's architecture and services," M.S. thesis, Riga Technical University, Riga, Latvia, 2010.

[7] P. Rudzajs and M. Kirikova, "IT Knowledge Requirements Identification In Organizational Networks : Cooperation Between Industrial Organizations And Universities," in Proceedings of the 18th International Conference on Information Systems Development (ISD 2009), Nanchang, China: Springer, 2010, pp. 187-199.

[8] R. Strazdina, M. Kirikova, L. Penicina, and P. Rudzajs, "Knowledge Requirements Monitoring System: Advantages for Industry and University," in 2010 Second International Conference on Information, Process, and Knowledge Management, Ieee, 2010, pp. 120-125.

[9] R. Strazdina, M. Kirikova, and P. Rudzajs, "Knowledge Integration Points in Contemporary Business Informatics," in Proceeding of the 9th International Conference on Perspectives in Business Informatics Research (BIR 2010), Rostock, Germany: 2010, pp. 33-42

[10] European Committee for Standardization, "European e-Competence Framework 1.0," 2008.

[11] SFIA FOUNDATION, "Framework reference SFIA version 4," 2008.

[12] European Committee for Standardization, "User guidelines for the application of the European e-Competence Framework," CEN, 2008.

[13] Association for Computing Machinery “CS 2008 Body of Knowledge”, 2010, Available: http://www.acm.org/education/curricula/ComputerScience2008.pdf. [Accessed: Dec. 9, 2010].

[14] M. Kirikova, R. Strazdina, I. Andersone, and U. Sukovskis, "Quality of Study Programs: An Ecosystems Perspective," in Proceedings of the Advances in Databases and Information Systems, Associated Workshops and Doctoral Consortium of the 13th East European Conference (ADBIS2009), Riga, Latvia: Springer Berlin/Heidelberg, 2010, pp. 39-46.

[15] R. Nikolov, "A Model for European e-Competence Framework Development in a University Environment," Stimulating Personal Development and Knowledge Sharing, 2008.

[16] C. Sanin, E. Szczerbicki, and C. Toro, "Combining Technologies To Achieve Decisional Trust," Cybernetics and Systems, vol. 39, no. 7, pp. 743-752, Oct.2008.

[17] F. Lin, J. Butters, K. Sandkuhl, and F. Ciravegna, "Context-based Ontology Matching: Concept and Application Cases," in 2010 10th IEEE International Conference on Computer and Information Technology, Ieee, 2010, pp. 1292-1298.

[18] J. Ć. Euzenat and P. Shvaiko, Ontology Matching. New York: Springer-Verlag Inc, 2007, p. 333.

[19] K. Janowicz and C. Kessler, "The role of ontology in improving gazetteer interaction," International Journal of Geographical Information Science, vol. 22, no. 10, pp. 1129-1157, 2008.

[20] Onet Center, "Onet resource center," 2010. Available: http://www.onetcenter.org/supplemental.html. [Accessed: Sept. 25, 2010].

[21] E. Schallehn, I. Geist, and K. U. Sattler, "Supporting Similarity Operations Based on Approximate String Matching on the Web," in On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, 3290 ed. R. Meersman and Z. Tari, Eds. Springer Berlin / Heidelberg, 2004, pp. 227-244.

Commonsense Knowledge Acquisition Using Compositional Relational Semantics

Hakki C. Cankaya
Dept. of Computer Engineering
Izmir Univ. of Economics, Izmir, Turkey
Email: hakki.cankaya@ieu.edu.tr

Eduardo Blanco and Dan Moldovan
Dept. of Computer Science
Univ. of Texas at Dallas, Richardson, Texas
Email: {eduardo, moldovan}@hlt.utdallas.edu

***Abstract*—A method for the acquisition of commonsense knowledge based on instantiations of metarules is presented. The metarules refer to some properties and objects that have those properties. Metarules are instantiated by automatically identifying objects that have those properties. In order to increase the applicability of a commonsense property to objects, composition of semantic relations is used. The method has been implemented and tested over WordNet. Results show that a commonsense metarule can produce many knowledge base axioms.**

***Keywords*-knowledge acquisition; commonsense knowledge; semantics.**

I. Introduction

Commonsense knowledge encompasses information people use everyday and it is assumed known by an average person; thus, it is not communicated most of the time. This makes more difficult to automate the acquisition of commonsense knowledge. To alleviate the problem of automatically extracting commonsense knowledge, semiautomatic approaches have been studied, where the system is given some seed information and is expected to generate more knowledge. There have been proposals to acquire commonsense knowledge from different sources by using different techniques. Some used collaborative efforts of experts and general public over the Web [1], [2]. There are other similar distributed human projects to collect commonsense knowledge [3]. Some proposals link the information obtained by the collaborative effort to known ontologies to expand and structure the commonsense knowledge [4], [5]. Some other proposals used text and World Wide Web as the source for commonsense knowledge acquisition [6], [7]. Despite these and other attempts, there is still a need for developing robust methods for automatic commonsense knowledge acquisition. In this paper, we introduce a new method for extracting commonsense knowledge by using metarules that contain user given commonsense rules and semantic relations.

II. Approach

The approach for extracting commonsense knowledge is based on metarules that contain commonsense rules provided by the user. These are then instantiated on a lexical knowledge base to identify large number of objects to which a high level commonsense rule applies. Commonsense rules refer to some common properties well known by average people. For example one can *see-thru* objects that have the *transparency* property.

To infer more commonsense knowledge of this type, the method automatically identifies in WordNet, or any other lexical source, the objects that have a property by searching for certain semantic relations. For example the object *glass* has the transparency property encoded by semantic relations in a lexical database. The inference mechanism used in the method concludes that one can see-thru glass since it has the transparency property. These instantiations of commonsense rules generate commonsense knowledge axioms.

The proposed method can accommodate potential restrictions and exceptions of a given commonsense rule. For example, some types of glass, like opaque glass, have to be excluded from the commonsense rule as they may not be *see-thru*. In order to find more objects that display a property the method searches for hyponyms of the objects that possess a given property since these also inherit that property, unless there is an exception. For example, the method extracts *round glass* as an object one can *see-thru*.

The mechanism for linking an object with other objects that have the same property relies on composition of semantic relations. The same mechanism is used to expand commonsense rules by cause and goal semantic relations. For example, *see-thru* causes more objects to be visible.

The method offers different metarules, because there are cases where semantic gaps cannot be bridged by the composition of semantic relations. For example, cars have *windshields* that are *transparent*. Even though cars are not *transparent*, one can *see-thru* a car. So, some objects don't inherit the property from their parts by using a Part-Whole relation in composition of semantic relations; however they inherit the rule and generate commonsense knowledge of the same type. The method can simply bridge those semantic gaps by using different metarules.

III. Semantic Relations and Compositional Relational Semantics

A. Semantic Relations

Semantic relations are the underlying relations between concepts expressed by words. They are implicit associations between chunks of text. Formally, a semantic relation is

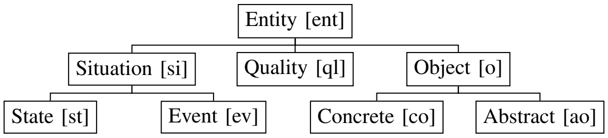


Figure 1. The ontology of sorts used to define DOMAIN(R) and RANGE(R).

represented as R(x, y), where R is the relation type, x the first argument and y the second. R(x, y) should be read as *x is* R *of y*, e.g., ISA(*gas guzzler*, *car*) should be read *gas guzzler* ISA *car*. The inverse of the relation is defined by R^{-1}(x, y), which is equal to R(y, x). Given R, we can define DOMAIN(R) and RANGE(R) as the set of sorts of concepts that can be part of the first and second argument, respectively. R(x, y) is formally defined by stating: *a*) relation type R, *b*) DOMAIN(R); and *c*) RANGE(R).

In order to define DOMAIN(R) and RANGE(R), we use the ontology depicted in Figure 1, which is a reduced version of [8]. The root corresponds to `entities`, which refer to all things about which something can be said. `Situation` is anything that happens at a time and place. Simply put, if one can think of the time and location of an entity, it is a `situation`. If they change the status of other entities, they are called `events` (e.g. *mix*, *grow*), otherwise `states` (e.g. *be standing next to the door*, *account for 10% of the sales*). `Objects` can be either `concrete` or `abstract`. The former occupy space, are touchable, tangible (e.g. *John*, *car*). The later are intangible, they are somehow product of human reasoning (*thought*, *music*). `Qualities` represent characteristics that can be assigned to entities, e.g., *tall*, *heavy*.

In this work, we use a particular set of five relations that are useful for commonsense extraction. This set, depicted in Table I, does not encode all the semantics in a text by any means. However, these relations help inferring commonsense knowledge as shown in the next section.

REASON(x, y) [REA] is defined as a broad relation in which x has a direct impact on y[1], eg, *[They don't smoke]$_y$ because [it is forbidden]$_x$*. GOAL(x, y) [GOA] encodes intentions, purposes, plans and intended consequences, e.g. *[Half of the garage]$_y$ is used for [storage]$_x$*. [PRO] captures the fact that x is a characteristic, property or value for y, eg, PRO(*tall*, *John*). PART-WHOLE(x, y) [PW] encodes the meronymy relation, i.e., x is a constituent part or a member of y. For example, PW(*engine*, *car*). ISA(x, y) [ISA] encodes x is a specialization of y, e.g., ISA(*adult*, *human*).

B. Commonsense Rules as Pseudo Relations

Sometimes it is useful to define and treat a particular connection between two entities like a semantic relation even though it is not one in the pure sense. By doing so we can use the formal framework of Compositional Relational Semantics and combine it with any other given semantic relations. We call this kind of connection *pseudo relation*, since they are not pure relations but are treated as such.

Table I
THE SET OF SEMANTIC RELATIONS USED IN THIS PAPER

Relation	Abbr.	DOMAIN × RANGE
REASON	REA	[si]×[si]
GOAL	GOA	[si ∪ ao]×[si ∪ o]
PROPERTY	PRO	[ql]×[o]
PART-WHOLE	PW	[o]×[o]
ISA	ISA	[o]×[o]

In this work, we define the pseudo relation COMMONSENSE_RULE (CS_R). CS_R(r, p) defines a connection between a situation r that applies given a certain property p. The connection has to have a commonsense nature, meaning that it is rarely explicitly stated. The complete definition is DOMAIN(CS_R) = [si], RANGE(CS_R) = [ql].

C. Compositional Relational Semantics (CRS)

The goal of composing, or linking semantic relations is to acquire new semantic relations by applying inference rules over already identified relations. An inference rule takes as input a set of semantic relations, called premises, and yields a conclusion. We define an inference rule by using the composition operator (∘). Formally, $\text{R}_1(x, y) \circ \text{R}_2(y, z) \rightarrow \text{R}_3(x, z)$, where R_1 and R_2 are the premises and R_3 is the conclusion.

In order to apply the composition operator over R_1 and R_2 they must fulfill the following necessary conditions: (a) R_1 and R_2 must be compatible; and (b) the second argument of R_1 and the first of R_2 must be the same concept, y.

a) Two relations R_1 and R_2 are compatible iff $\text{RANGE}(\text{R}_1) \cap \text{DOMAIN}(\text{R}_2) \neq \emptyset$. Say, we have an inference rule, PRO(p, x) ∘ISA^{-1}(x, y) → PRO(p, y), which means that if p is a property of x and x is the hypernym of y, then y inherits the property p. This inference rule actually holds because PRO and ISA^{-1} are compatible in this case, RANGE(PRO) ∩ DOMAIN(ISA^{-1}) = [o].

b) In an instance of the inference rule above, PRO(*sharpness*, *knife*) ∘ISA^{-1}(*knife*, *butcher-knife*) → PRO(*sharpness*, *butcher-knife*), there is a common concept *knife* that links the premises of the inference rule, which fulfills the second requirement of the compositional relational semantics. The conclusion is if *knife* has property *sharpness*, then any hyponym of *knife*, like *butcher-knife* inherits the property unless stated otherwise.

IV. METHOD

The proposed method for commonsense extraction follows a semiautomatic approach. Given a commonsense rule that applies to a certain property, the method uses metarules in order to extract commonsense knowledge. The method exploits properties of objects, the rules that apply to them

[1] It includes relations usually named CAUSE and INFLUENCE.

and how they can be transferred thru a chain of semantic relations. Extensions to the method have been studied to automatically infer more properties and commonsense rules, significantly increasing the amount of knowledge extracted. All the inferences are performed within the framework of Compositional Relational Semantics.

A. Metarules

Two main metarules are used to obtain commonsense knowledge.

1) Metarule 1: $\text{CS\_R}(r,p) \circ \text{PRO}(p,x) \rightarrow \text{CS}(r,x)$.

Rationale: *rule r applies to property p; p is a property of x; therefore, r applies to x.*

Example: Given the commonsense rule *you cannot check in for flight sharp objects*, CS_R(*cannot check in for flight*, *sharp*), and the fact that knifes are sharp, PRO(*sharp*, *knife*), we obtain the commonsense knowledge that knifes cannot be checked in for flight, CS(*cannot check in for flight*, *knife*). Formally, $\text{CS\_R}(\textit{cannot check in for flight}, \textit{sharp}) \circ \text{PRO}(\textit{sharp}, \textit{knife}) \rightarrow \text{CS}(\textit{cannot check in for flight}, \textit{knife})$.

The columns *rule(r), property (p), and concepts (x)* in Table II shows examples of knowledge extracted using this metarule.

Some objects x are parts or members of larger objects y. Metarule 1 can be expanded by adding a part-whole relation to the premise, resulting in a new metarule, Metarule 2.

2) Metarule 2: $\text{CS\_R}(r,p) \circ \text{PRO}(p,x) \circ \text{PW}(x,y) \rightarrow \text{CS}(r,y)$

Rationale: *rule r applies to property p; p is a property of x; x is a part of y; therefore, r applies to y.*

Example: Given the commonsense rule *electric objects need power to operate*, CS_R(*need power*, *electric*), electric is a property of electric motors PRO(*electric*, *motor*), the fact that electric motors are components of electric fans, PW(*motor*, *fan*), we obtain CS(*need power*, *fan*), i.e., the commonsense knowledge that fans need power to operate. Formally, $\text{CS\_R}(\textit{need power}, \textit{electric}) \circ \text{PRO}(\textit{electric}, \textit{motor}) \circ \text{PW}(\textit{motor}, \textit{fan}) \rightarrow \text{CS}(\textit{need power}, \textit{fan})$.

B. Restrictions and Exceptions

The metarules introduced so far do not have any restrictions on the kind of concepts they link. However, a closer inspection leads to the conclusion that sometimes restrictions and exceptions have to be imposed in order to guarantee a high accuracy in the inferences performed. Restrictions and exceptions are indicated between brackets and added at the end of the premises with an & operator. Formally, we denote restrictions for an axiom as $\text{R}_1(x,y) \circ \text{R}_2(y,z) \& [\textit{restrictions}] \rightarrow \text{R}_3(x,z)$. An axiom performs an inference only if all the restrictions are fulfilled.

For example, something portable can be carried, but constraints on the weight and the person carrying the object are necessary. A child can carry a *watch*, but will have trouble carrying a *portable television set*. Consider the commonsense rule *eating sweets excessively results in weight gain*. An exception to this rule is *saccharin* which is sweet but *calorie-free*. Thus, an exception is attached to the rule.

The Metarule 2 makes the wholes inherit the rules that apply to the properties of its parts. Several restrictions should be placed in order to avoid invalid inferences.

First, *r* should not describe any physical property such as weight or size. One can *lift light objects*, and *car seat cushions* are *light* and part of *cars*, and yet one cannot lift cars. In other words, rules that state physical properties of parts do not transfer to their wholes.

Second, *r* should not encode an `event` (`ev`). Following Table II, only the rules encoding a `state` (`st`) can be used with Metarule 2. For example, *one can not see alive animals that are extinct* CS_R(*[cannot see alive]$_{st}$*, *extinct.j.1*). Since it encodes a state, the wholes inherit the rule: if *y* has a part *x* which is extinct, one cannot see *y* alive. On the other hand, consider CS_R(*[will roll on inclined path]$_{ev}$*, *round.j.1*). Just because *y* (mouse.n.2) has a round part *x* (ball.n.3), *y* will not roll on an inclined path. Similarly, a *removable cup holder* (*x*) is *portable* (*p*) and part of a *car* (*y*), and one *[can carry]$_{ev}$* (*r*) portable objects. Because *r* is an `event`, we cannot instantiate Metarule 2 and infer CS(*can carry with you*, *car*).

Formally, the final definition of both metarules are:

- $\text{CS\_R}(x,y) \circ \text{PRO}(y,z) \& [\textit{rest(x)}] \rightarrow \text{CS}(x,z)$
- $\text{CS\_R}(r,p) \circ \text{PRO}(p,x) \circ \text{PW}(x,y) \& [\textit{r is a st, no physical properties}] \rightarrow \text{CS\_R}(r,y)$.

C. Extensions using Compositional Relational Semantics

In this section, we aim to automatically extend the commonsense rules (CS_R) and object properties (PRO) by chaining semantic relations. The result is more inferences performed by both metarules and therefore more commonsense knowledge is extracted. We do so by combining CS_R and PRO with semantic relations and the rules of compositional semantics.

1) Rule Extension: Given a rule *r* that applies to a certain property *p*, one can also infer that *a*) actions whose goal is to achieve *p*; and *b*) the goals and effects of *r* also apply to *p*. Formally $\text{CS\_R}(x,y) \circ \text{GOA}(y,z) \rightarrow \text{CS\_R}(x,z)$, $\text{CS\_R}(x,y) \circ \text{GOA}^{-1}(y,z) \rightarrow \text{CS\_R}(x,z)$ and $\text{REA}^{-1}(x,y) \circ \text{CS\_R}(y,z) \rightarrow \text{CS\_R}(x,z)$.

For example, given CS_R(*can-be-seen*, *visible*), and knowing that one foregrounds (foregrd) in order to make visible (GOA(*visible*, *foregrd*)), we obtain CS_R(*can-be-seen*, *foregrd*). Formally, $\text{CS\_R}(\textit{can-be-seen}, \textit{visible}) \circ \text{GOA}(\textit{visible}, \textit{foregrd}) \rightarrow \text{CS\_R}(\textit{can-be-seen}, \textit{foregrd})$. Similarly, as seen in Table II given CS_R(*spills if not in container*, *liquid.j.1*), and knowing that something flows if spilled, REA^{-1}(*flow*, *spill*), we obtain CS_R(*flow if not in container*, *liquid.j.1*).

Table II
EXAMPLES OF KNOWLEDGE EXTRACTED USING THE METARULE 1, $\text{CS\_R}(r, p) \circ \text{PRO}(p, x) \rightarrow \text{CS}(r, x)$ AND EXTENSIONS.

sort	*rule extension*	*rule (r)*	*property (p)*	*concept (x)*	*property extension*
st	cannot be blind	can see thru	transparent.j.1	window.n.1, lens.n.1	rear_window.n.1, quarterlight.n.1, contact_lens.n.1, condenser.n.4
	-	cannot check in for flight	sharp.j.1	knife.n.1, parer.n.2	slicer.n.3, carving_knife.n.1
	-	cannot see alive	extinct.j.1	dinosaur.n.1, moa.n.1	trachodon.n.1, ornithomimid.n.1, anomalopteryx.n.1
	-	cannot touch	imaginary.j.1	bogeyman.n.1, equator.n.1	-
	-	not likable	annoying.j.1	pest.n.1, trial.n.6	nudnik.n.1
ev	-	excess results in weight gain	sweet.j.1	jimmies.n.1, muffin.n.1	popover.n.1, corn_muffin.n.1
	-	you can carry with you	portable.j.1	watch.n.1, flashlight.n.1	pocket_watch.n.1, digital_watch.n.1, penlight.n.1
	will move on inclined path	will roll on inclined path	round.j.1	ball.n.1	golf_ball.n.1, polo_ball.n.1
	flows if not in a container	spills if not in container	liquid.j.1	beverage.n.1, soup.n.1, draft.n.8	softdrink.n.1, coke.n.1, potage.n.1, gazpacho.n.1, vichyssoise.n.1
	can cater, can cook	can eat / consume	edible.j.1	potato.n.1, radish.n.1	french_fires.n.1, mashed_potato.n.1

2) Property Extension: Given the fact that a certain p is a property of x, one can also infer that all hyponyms (ISA^{-1}) of x have that property. Formally, $\text{PRO}(x, y) \circ \text{ISA}^{-1}(y, z) \rightarrow \text{PRO}(x, z)$. For example, in Table II given PRO(*liquid.j.1*, *beverage.n.1*), and knowing that ISA^{-1}(*beverage.n.1*, *soft drink.n.1*) and ISA^{-1}(*beverage.n.1*, *coke.n.1*), we obtain PRO(*liquid.j.1*, *soft drink.n.1*) and PRO(*liquid.j.1*, *coke.n.1*).

One might be tempted to follow the intuition that wholes inherit the properties of its parts. However, closer inspection reveals that this plausible axiom does not hold: cars have as parts windows, windows are transparent, and yet cars are not transparent.

V. IMPLEMENTATION AND RESULTS

In order to automatically instantiate the metarules to identify objects x that have properties p and benefit from the power of the method described in Section 4, it is necessary to have semantic relations readily available. In our experiments, the commonsense rules were provided by humans, including corresponding restrictions and exceptions. For the semantic relations that are necessary for instantiations and extensions, we used an annotated resource called eXtended WordNet-Knowledge Base (XWN-KB).

A. eXtended WordNet Knowledge Base (XWN-KB)

Table III
XWN-KB REPRESENTATION OF A CONCEPT

Knife#2: a weapon with a handle and blade with a sharp point	
ISA(*knife, weapon*)	PW(*handle, knife*)
PW(*blade, knife*)	PRO(*sharp, knife*)

The XWN-KB is an upper ontology built as an extension to eXtended WordNet (XWN) which is derived from WordNet (WN) [9]. The novelty that XWN-KB offers is that the glosses of synsets have been transformed into semantic relations by using a reliable semantic parser and partly verified by human annotators. The result is a knowledge base that is highly interconnected. Unlike a domain specific ontology that is narrow, the XWN-KB uses definitional glosses of WordNet synsets which are regarded as universal knowledge. WordNet and its extensions offer a large and reliable world knowledge source for extracting commonsense knowledge by applying metarules.

For example, the WordNet concept *knife* in sense #2 has the following gloss: *a weapon with a handle and blade with a sharp point.* In XWN-KB this text definition has been transformed into a set of semantic relations as shown in Table III. For us important are PRO(*sharp*, *knife*), PW(*handle*, *knife*), and ISA(*knife*, *weapon*). When CS-R(cannot check in for flight, sharp) is given, the method searches for a property relation and this instantiates the concept *knife#2* by locating its PRO(*sharp*, *knife*) relation. For extensions, it uses the mechanism of composition of semantic relations over the annotated semantic relations provided by the XWN-KB.

B. Implementation

The implementation is coded by perl and python scripts that interface with XWN and WN. The set of commonsense rules are given to the code. The code applies metarules to the given commonsense rules following the procedure below.

Input: A set of commonsense rules.

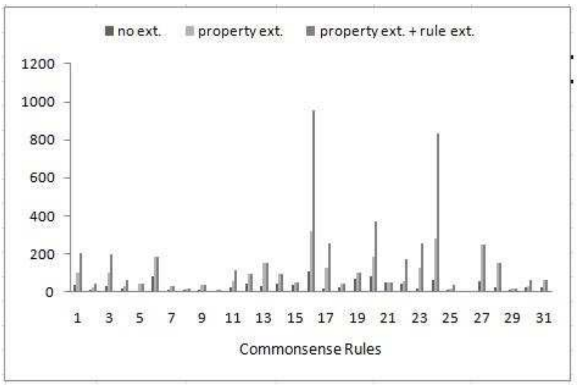


Figure 2. Results for Metarule1

Output: Collection of commonsense axioms for all given rules: $S_x[]$ + $S_y[]$.

Main-Procedure: For each commonsense rule, repeat the steps below:

1. Apply Metarule 1 to the commonsense rule. Instantiate all concepts $c_x[]$ that have the property given in the commonsense rule.

1.1. For each concept in $c_x[]$, process the property extension and find all hyponyms, $h_x[]$. Accumulate all $h_x[]$,

$H_x[] \leftarrow H_x[] + h_x[]$

1.2. Process rule extension for the commonsense rule and calculate all other rules, $k_x[]$. Apply the new rules to all concepts and store the final commonsense axioms,

$S_x[] \leftarrow S_x[] + \{H_x[] + c_x[]\} \times k_x[]$

2. Apply Metarule 2 to all $c_x[]$. Instantiate all concepts $c_y[]$ that inherit the rule.

2.1. Apply Metarule 2 to all $H_x[]$ (calculated previously) and find all concepts $h_y[]$ that inherit the rule. Accumulate all $h_y[]$,

$H_y[] \leftarrow H_y[] + h_y[]$

2.2. Process the rule extension and calculate all other rules $k_y[]$ for Metarule 2. Apply the new rules to all concepts and store the final commonsense axioms,

$S_y[] \leftarrow S_y[] + \{H_y[] + c_y[]\} \times k_y[]$

C. Results on XWN-KB

Following the procedure, a set of 32 commonsense rules was provided as input to the implementation. Metarule 1 has instantiated 1015 commonsense axioms for the given set without any extensions (see Table IV). Then, the property extension was performed by using composition of semantic relations and 2833 axioms were generated this way. Human validation was performed and only 46 generated axioms were tagged as incorrect yielding a precision of 0.984. As explained in earlier sections, the rule extension augments the commonsense rule that applies to the property. All new rules that are generated by the rule extension can also apply to all property inheriting objects including those that are generated by the property extension. Therefore, the cardinality of this rule augmentation becomes a multiplying factor. For example, for a commonsense rule i, CS-$R_i(p_i,r_i)$, the number of concepts that are instantiated and applied to the rule is S_i and the number of extra objects that are found by the property extension is L_i. If the number of new rules that are generated by the rule extension is R_i, the total number of generated axioms is $T = \sum_i (S_i + L_i) * R_i$. Therefore, the rule extension is a rather powerful factor. In the implementation, for Metarule 1, the total number of axioms is 4938 (see Table IV). Figure 2 plots for all rules S_i (no ext.), S_i+L_i (property ext.), and T_i (property ext. + rule ext.) values. The observation of the results reveals that there is quite some variation among the commonsense rules in terms of their S_i, S_i+ L_i, and T_i values. The variation in S_i is caused by the frequency of the property and is related with the number of concepts in the knowledge base that in fact has the property in its gloss and semantic relations. The L_i depends on the hyponmym connectivity of the concept that has the property. Basically more hyponyms result in larger L_i value for rule i.

We also looked at precision values for all 32 commonsense rules and compared them in Figure 3 with and without extensions for Metarule 1. In the experiment, while extensions increased the generated commonsense axioms significantly, the precision did not deteriorate. However, this purely depends on the resource used. And the propagation of errors depends on the hyponym connectivity of the concepts. For example, if the incorrect concepts that are initiated by the metarule have high hyponym connectivity, then the chances are high for obtaining a poor precision, since the error has the ripple effect. So, the authors' suggestion for potential implementations of the method is to introduce an annotation step between the metarule instantiations and the extensions, so that incorrect concepts are weeded out before they ripple and adversely affect the performance.

Experiments showed that with the given set of commonsense rules, Metarules work differently. Even though there were some instantiations where rule applied to the concept without inheriting the property. However, those few results were augmented by the rule extension, increasing the final count for the commonsense knowledge.

The results of the numerical study seems promissing. Starting with 32 commonsense rules provided by the user, the method generated 4950 commonsense axioms, more than two orders of magnitude increase.

VI. Applications

Commonsense knowledge can be used in many applications that require some form of reasoning. It is used to bridge knowledge gaps and leads to solutions which may not be possible otherwise. Such applications are question answering (Q/A), text entailment systems (RTE), search engines, multi agent systems, etc. In question answering systems, the commonsense knowledge can play a significant role in answering questions that seem trivial for humans but are nearly impossible for machines. Below is an example

Table IV
NUMERICAL RESULTS

	No Extension		Extension 1		Extension 2		Extension 1 & 2	
	Conclusion	**Precision**	**Conclusion**	**Precision**	**Conclusion**	**Precision**	**Conclusion**	**Precision**
Metarule 1	1015	0.974	2833	0.984	1696	0.972	4938	0.985
Metarule 2	7	0.714	7	0.714	12	0.666	12	0.666
Total	1022	0.972	2840	0.983	1708	0.970	4950	0.984

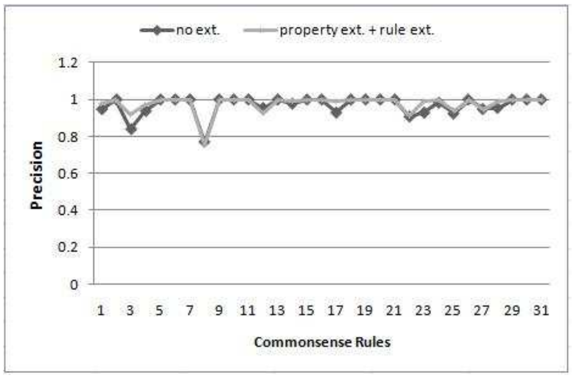


Figure 3. Precision for Metarule 1

from TREC2007. Even though the system used was a high performance system, it could not compute an answer for the question:

Question, Q2.21600004: (Paul Krugman) What is Krugman's academic specialty?

Answer: Economics

Text in BLOG06-20051213-068-0019517474: "..Paul Krugman is Professor of Economics at Princeton University,....."

To answer this question a connection has to be made between *academic* and *economics*. This cannot be done using basic lexical chains in WordNet alone. An axiom establishing this connection was generated by using the proposed method. We run the program of the proposed method with a rule that has *academic* as the property and received a list of concepts that have this property in XWN-KB. One of the concepts in the list is *economics-department* which claimed *academic* as an inherited property. This easily bridges the semantic gap to reach the answer stating that *Krugman is Professor of Economics in Princeton.*

VII. CONCLUSIONS

The resource used in this paper is eXtended WordNet Knowledge Base, simply because it has already synset glosses transformed into semantic relations and it contains information that is widely applicable. Any corpora, including the Internet, that is semantically parsed and transformed into semantic relations can be used in our method. The accuracy of the results is highly correlated to the accuracy of the semantic relations extracted from text.

The method presented here has the disadvantage that users need to provide commonsense rules that are then automatically instantiated to a large number of objects. We found that an average person can come up rather quickly with many commonsense rules but it is nearly impossible for humans to quickly think of many possible instantiations of these rules. In this sense the method introduced here automates the most difficult part of commonsense knowledge acquisition. The method proposes entensions of metarules that rely on compositional relational semantics a powerful technique to increase its generative power.

REFERENCES

[1] D. B. Lenat, "Cyc: Large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 28–32, 1995.

[2] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems.* Lecture Notes in Computer Science, No 2519, Heidelberg, Springer-Verlag, 2002.

[3] L. V. Ahn, M. Kedia, and M. Blum, "Verbosity: A game for collecting common-sense knowledge," in *ACM Conference on Human Factors Computing System*, 2006, pp. 75–78.

[4] H. Liu and P. Singh, "Conceptnet: a practical commonsense reasoning toot-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.

[5] F. Gomez, "The acquisition of common sense knowledge by being told: An application of nlp to itself." Lecture Notes in Computer Science, Vol. 5039 2519, Heidelberg, Springer-Verlag, 2008.

[6] V. C. Storey, S. V., and Y. Ding, "A semi-automatic approach to extracting common sense knowledge from knowledge sources." Lecture Notes in Computer Science, Vol. 3513, Heidelberg, Springer-Verlag, 2005.

[7] Y. Zhui, L.-J. Zhan, D.-S. Wang, and C.-G. Cao, "Manual experiment on commonsense knowledge acquisition from web corpora," in *Proceedings of International Conference on Machine Learning and Cybernetics*, 2008.

[8] H. Helbig, *Knowledge Representation and the Semantics of Natural Language*, 1st ed. Springer, 2005.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database and Some of its Applications.* The MIT Press, 1998.

About an Architecture for Integrated Content-Based Enterprise Search

Manfred Grauer, Ulf Müller, Daniel Metz, Sachin S. Karadgi, Walter Schäfer
Information Systems Institute, University of Siegen, Germany
Email: {grauer, mueller, metz, karadgi, jonas}@fb5.uni-siegen.de

***Abstract* – The main goal of knowledge management is to improve the management of information and knowledge within and across enterprises. Enterprise knowledge is embedded in enterprise's data managed in a wide range of information systems (e.g., product data management, enterprise resource planning systems). These enterprise data can correspond to textual, numerical or multimedia. In the past, several search systems have been developed to provide access to primarily text-based enterprise content like web pages, data stored in database systems or emails. Similarly, dedicated search systems exist for searching information from multimedia data. As these search systems focus on particular enterprise content (e.g., textual data), they lack in providing a holistic view on available enterprise's knowledge. Therefore in the current contribution, architecture for integrated content-based enterprise search encompassing various enterprise data sources is elaborated. This architecture supports the exploitation of embedded enterprise knowledge. Further, the architecture is validated in an industrial scenario.**

Keywords* – *enterprise search; knowledge management; information retrieval; shape matching; image-based retrieval.

I. Introduction

Alaavi and Leidner define "data as raw numbers and facts, information as processed data, and knowledge as authenticated information" [1]. Knowledge is embedded into enterprise processes by enterprise members and described through various data types as illustrated in Fig. 1 (e.g., documents, 3D models, emails). This knowledge is enriched and enlarged as the enterprise process moves from upstream to downstream processes. Hence, necessitates for organizing and managing enterprise data in a manner that it can be identified quickly and assimilated by enterprise members for design and execution of enterprise processes.

In engineering, 3D computer-aided design (CAD) models, 2D drawings and design patterns have to be created during product development process [2]. Information and library sciences are managing text-based documents like books, journals and magazines. Order information and its corresponding offer documents, bill of material (BOM), email interaction and minutes of meeting have to be managed by sales department. Patient records have to be managed by hospitals. Due to the aforementioned diversity of applications, specialized IT-systems have been developed to manage different types of enterprise data (e.g., enterprise resource planning (ERP), customer relationship management (CRM), product data management (PDM) or clinical information systems (CIS)). Each of these IT-systems contains a large amount of enterprise data related to different enterprise entity types (e.g., products, customers, patients).

Many of the enterprises utilize aforesaid IT-systems in various combinations. Therefore, the IT landscape of an enterprise is often scattered and inadequately integrated i.e., enterprises often need to address data interoperability issues. As a consequence, enterprise members have to perform multiple searches for information and corresponding knowledge related to an enterprise entity in various IT-systems. In majority of the research, search system functionality is achieved by means of organising and retrieving structured and unstructured textual data from a certain IT-system [3]. Also, multimedia data (e.g., images, videos, sounds) is rarely considered, and only special research areas deal with organizing and retrieving of multimedia objects [4][5].

Overall, lack of integrated data from different IT-systems hinders establishing holistic view about the requested enterprise entity. To overcome these restrictions and drawbacks, enterprise search (ES) has emerged as a promising research area. ES is the practice of making enterprise content from numerous IT-systems, such as databases and intranets, searchable to certain stakeholders [3][6]. The vision is that an enterprise member is capable to request all necessary information and associated knowledge with minimum effort from various IT-systems, which is required to effectively design and execute enterprise processes. Therefore, an ES engine has to be provided to enterprise member incorporating enterprise data from different IT-systems.

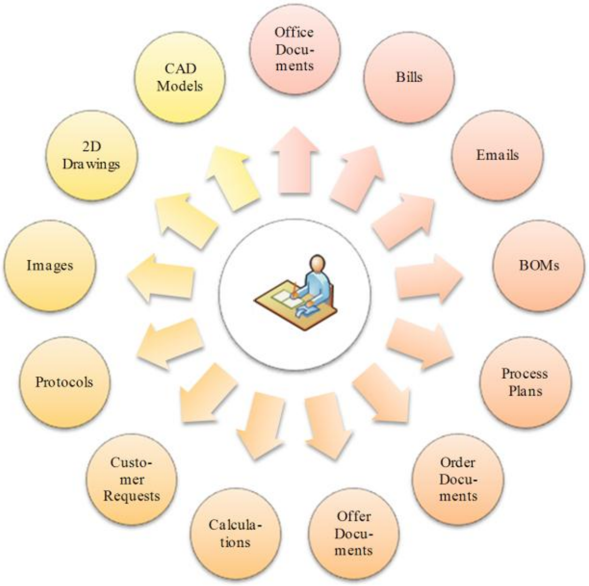


Figure 1. Different data sources and document types available to enterprise members

To address the aforementioned requirements of ES, architecture for integrated content-based enterprise search (CBES) is presented. The contribution is structured as follows. An overview of current research trends in ES and content-based retrieval is presented in section II. Section III motivates CBES, and defines challenges and requirements associated with CBES. The envisaged architecture and its components are elaborated in section IV. To validate the envisaged architecture, Section V describes an industrial case study. Finally, conclusion is presented in Section VI.

II. RELATED WORK

The main task of ES is to improve information and knowledge management, and facilitate information access within and across enterprises [7]. ES allows enterprise members to search through documents, emails and other data sources available in an enterprise to identify information and associated knowledge [7]. Enterprise members spend one third of their work time searching for information necessary for designing and executing enterprise processes [8]. Therefore, it is not surprising that the investment in development of comprehensive ES during the last years has increased and major companies like Google, Microsoft, IBM or SAP have developed specific and proprietary ES-systems [9][10].

Data in an enterprise is primarily textual, in the form of (intranet) web pages, emails, orders details, bills, reports, and so forth (s. Fig. 1). Hence, majority of the ES implementations are focused on textual data [3][4][8]. An early approach of ES has been named as an enterprise intelligent system [11]. This system tries to integrate data from intranet servers, web servers and web search services. The usage of different search techniques (e.g., keyword search, semantic search) were analysed, and utilized in search of information from integrated vehicle health management data [12]. Similarly, ontologies were used to overcome the limitations of textual retrieval [13]. Further, an engineering information retrieval system has been elaborated which attempts to integrate data from various engineering data sources into a centralized dataset and make it accessible via ontology-based retrieval [14].

A search system for identifying information from huge collection of documents from digital libraries was presented [15]. The effectiveness of this search system was increased by exploiting techniques like full text search, collaborative filtering and multifaceted browsing. A search architecture was elaborated which employed the search functionalities of the source systems to locate relevant information [16]. Retrieved results are combined together in a global result list using case-based reasoning (CBR). Further, CBR was used for ranking the results in a result list.

Apart from textual data, nowadays multimedia data is widely used in enterprises. Multimedia data include images, photos, videos and sound recordings, among other. For instance, meeting involving enterprise members from multiple sites are recorded. Most of the available search systems have not integrated multimedia data along with the textual data. However, research has been carried out in isolation to search various multimedia data.

A survey of image retrieval techniques and systems were presented [17]. Image retrieval techniques were classified into two groups: text-based and content-based retrieval. Both groups deal with the same issue to overcome the gap between what is really experienced by the enterprise member when viewing the content on screen and what is really stored (e.g., in a database). This challenge is not only relevant for image retrieval but also for other kinds of multimedia data retrieval.

Several IT-systems have been developed for 3D model retrieval based on textual annotations, content, features and ontologies, and so forth [14][18][19]. All the presented approaches show that it is possible to search 3D content in large repositories. Another field in the context of multimedia data retrieval is face recognition. This field is relevant for identification of faces displayed on an image or in a video, and it has become more important from (public) security point of view. The state-of-the-art of face recognition techniques was carried out [20]. Besides images and photos, retrieval for other multimedia data is available (e.g., video, sound). Retrieval systems for these kinds of data have been developed and encouraging results have been achieved. A summary of video retrieval techniques can be found in [21] and survey of sound retrieval techniques is presented in [22].

The review of the related work has shown that effective techniques for content-based retrieval exist and dedicated systems for specific scenarios have been developed. However, an architecture which allows the incorporation of all aforementioned enterprise data types from different data sources could not be identified.

III. CHALLENGES AND REQUIREMENTS FOR AN ARCHITECTURE FOR INTEGRATED CBES

As already mentioned, most data in an enterprise is text-based like reports, emails, order details, and bills (s. Fig. 1). This obviously raises the question: why is it necessary to consider multimedia data and integrate it into an ES-strategy? The main reason is the increasing relevance for utilizing multimedia data during enterprise process design and execution. For example, design engineers and architects use 3D models to store information about design and construction of products or buildings. Respectively, images and drawings are used to identify and describe products, and voice recordings and videos are used to protocol meetings.

In current search environments, collections of multimedia data are managed by assigning textual annotations (i.e., textual metadata) which can be employed for retrieval [4]. Usually, enterprise members from different departments contribute during design and execution of enterprise processes. Different annotations are assigned to the same multimedia data depending on the executed process step and enterprise members' background. Therefore, knowledge managed using aforementioned metadata-based techniques places limitations during ES. In contrary to this metadata approach and to ensure that relevant product-related information and knowledge (e.g., embedded in a bill or 3D models) can be retrieved from different departments of an enterprise, it is mandatory to integrate different annotations, attributes from various IT-systems as well as content from

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-121-2

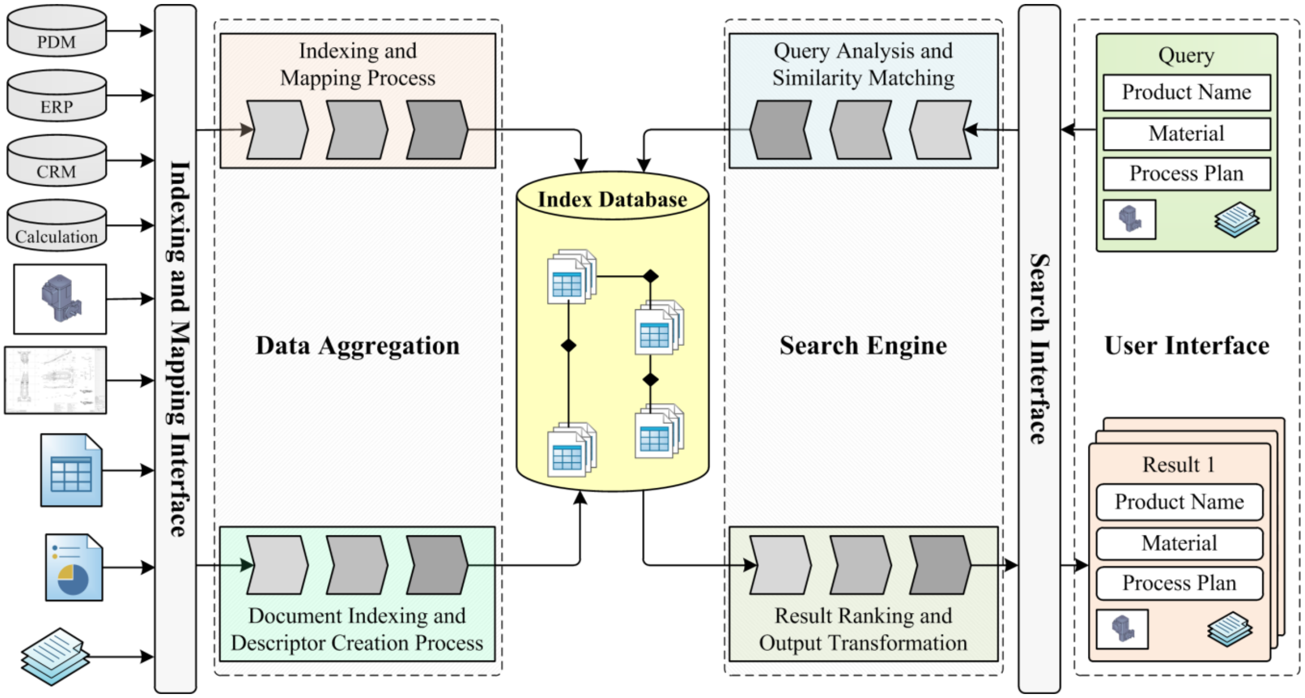

PDM
ERP
CRM
Calculation
Indexing and Mapping Interface
Indexing and Mapping Process
Data Aggregation
Document Indexing and Descriptor Creation Process
Index Database
Query Analysis and Similarity Matching
Search Engine
Result Ranking and Output Transformation
Search Interface
Query
Product Name
Material
Process Plan
User Interface
Result 1
Product Name
Material
Process Plan

Figure 2. Architecture of integrated content-based enterprise search (CBES)

multimedia data. The overall goal is to create an integrated descriptor for retrieving relevant enterprise process entities (e.g., orders, bills, 3D models). In addition, the envisaged ES has to fulfill certain requirements which can be classified into four groups: (i) coverage, (ii) security, (iii) query support, and (iv) presentation of results.

Coverage in the context of a holistic ES means that all relevant IT-systems and data sources are included into the enterprise search. The challenge is to integrate information and knowledge which is scattered across an enterprise. Besides information in IT-systems like CRM or ERP systems, information and knowledge can also be available in emails, text documents, 3D models and images, and so forth.

Security has gained considerable attention in the context of ES due to tightened legal requirements and enterprise's policies related to data privacy and confidentiality. Obviously, only privileged enterprise members should have access to enterprise data. An enterprise member should have access grants to relevant data, information and knowledge depending upon his position within a department and the activities he/she has to fulfill.

Apart from the scope of the ES and its security concepts, user-friendly and convenient search interfaces have to be provided to the enterprise members. The search interfaces have to respect an enterprise member's roles and privileges. For instance, an engineer will demand a dedicated search interface to search for 3D models where as sales department personnel will have a graphical user interface (GUI) to explore previous offers and orders. In either case, a search interface assists an enterprise member to input suitable number of search attributes. Based on these search attributes, appropriate similarity search techniques will be employed to construct the result list (e.g., containing 3D models or orders). In case of multimedia data, this can mean that functionality need to be provided where user can submit a template (e.g., 2D drawing of a product) to the search engine for requesting product related information and knowledge.

Last but not least, adequate presentation of the search result plays a prominent role. The retrieval system has to provide appropriate interfaces to display the result list, especially in case of multimedia data. For example, the ES-system provides a preview window for visualizing a 3D model of a selected result item. In addition, access to information and knowledge associated with the selected result item has to be implemented. Hence, the ES-system establishes links to the original data managed in one of the enterprise IT-systems.

IV. Architecture for Integrated CBES

As mentioned before, enterprise data is stored in different IT-systems (e.g., PDM system, ERP system). In addition, each data source has its own specialized data structures. For instance, a PDM system is used to manage product-related data created or modified along a product's life cycle [2]. The proposed process model for an integrated ES is composed of the following steps: (i) analysis of enterprise processes concerning their status and requirements regarding ES, (ii) create a centralized search index database with attributes and content-based descriptors revealed from enterprise data, (iii) engage search index to retrieve enterprise members' requested information, and (iv) evaluate, update and optimize the integrated ES system (i.e., especially the search index database).

Before implementing an ES system, it is mandatory to identify enterprise knowledge required by the enterprise members. Enterprise knowledge is used and shared by enterprise members along enterprise process execution. The knowledge flows within a certain enterprise process can be (re-)designed and analyzed employing the modeling and description language (KMDL) [23]. The identified characteristics of knowledge creation and assimilation assist the establishment of an appropriate search index database. In addition, the data sources required to execute enterprise processes have to be documented (e.g., with unified modeling language (UML)).

Based on the aforementioned (process) analysis, an appropriate index database has to be defined. This index database contains attributes and content-based descriptors related to enterprise process entities (e.g., orders, offers). Data mining techniques and structured interviews can be employed to determine relevant attributes and required content-based descriptors. The attributes and descriptors are derived from enterprise data which is managed in different enterprise IT-systems. Each entry in the index database points to the data sources.

Analysis of enterprise processes and setup of index database guides in definition of GUI and implementation of ES engine. GUI is utilized to submit search request to ES engine and display the search result list. Today’s enterprise processes are agile which requires periodical evaluation, refinement and optimization of index database and the corresponding ES-system. Therefore, enterprise process analysis has to be performed systematically to reflect the actual situation of an enterprise. Based on the envisaged process model for an integrated ES, architecture is elaborated in the following sub-sections.

The architectural overview of integrated CBES is illustrated in Fig. 2. The architecture is composed of four components: (i) data aggregation interfaces for indexing data sources and file types available in an enterprise, (ii) index database for textual and numerical attributes from the IT-systems and extract descriptors from multimedia data, (iii) ES engine to support integrated CBES, and (iv) GUI to the enterprise members based on their roles and privileges.

A. Data Aggregation

The data aggregation component provides functionalities to index and map enterprise data from different IT-systems, as depicted in Fig. 2. This component consists of two sub-components: (i) mapping and indexing of attributes from different IT-systems like ERP-, PDM- or CRM- systems to the columns of the predefined index database, and (ii) services for deriving content-based descriptors for multimedia data (e.g., images, 3D CAD models).

Mapping is used to establish a link between data fields of a source system (e.g., managed in a database or XML-structured files) to corresponding columns of the index database. Depending upon the data type of a source field (e.g., numerical, alphanumerical) dedicated indexing methods are employed to optimize the efficiency of the subsequent retrieval process. For instance, a text analysis and text aggregation component is available to extract the content from text documents. If a new value gets stored or an existing value gets updated in a certain source IT-system, the mapping functionality transfers this value to the appropriate column in the index database.

In the case of multimedia data, services are provided to extract relevant metadata from the corresponding file. The retrieved metadata or the calculated content-based descriptor is also stored in the index database. For each file type, a dedicated service is available which extracts the metadata and / or creates a content-based descriptor. For instance, a 3D model of a product is represented as a triangulated surface and stored as a stereolithography (STL) file. This file can be transformed into a content-based descriptor that describes the shape and dimensions of the product. This transformation process can be delineated in three steps (s. Fig. 3): (i) normalisation of the 3D shape essential to avoid problems resulting from translation, rotation and scaling invariance, (ii) creation of images from different perspectives of the normalized 3D model, and (iii) employ various algorithms on the taken images to create image-based descriptors (e.g., edge histogram).

B. Index Database

The index database contains an entry for each enterprise process entity of interest (e.g., order). The entries in the index database represent the enterprise process entities with a number of (alpha-) numerical attributes and if required, geometrical descriptors (s. Fig. 3). These attributes and descriptors are generated by the data aggregation component.

The stored data in the index database has to be indexed for efficient retrieval of enterprise process entities. Therefore, simple numerical and alphanumerical data types are indexed using the available indexing capabilities of relational databases. However, these indexing capabilities are not applicable for complex shape descriptors. Hence, feature space indexing structures (e.g., R*-tree) or metric indexing structures (e.g., M-tree) are available for indexing geometrical (shape-based) descriptors [24][25][26]. These indexing structures have been successfully used for retrieval of complex sheet metal components [27].

Obviously, index database and the available data sources

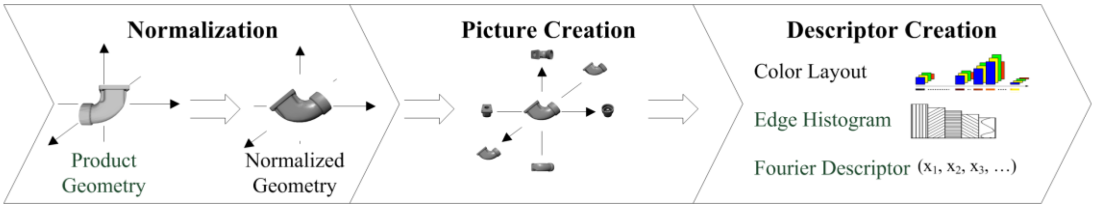


Figure 3. Illustration of image-based descriptor creation process for 3D models

of the enterprise's IT-systems (e.g., PDM system) have to be synchronised. Different methodologies are applicable to guarantee that data sources and index database are in synchronous state. Based on the enterprise's IT strategy, these methodologies can be classified as following: (i) update the index database just-in-time, i.e., when the original data in an enterprise's IT-system has been modified, (ii) employ a batch process to update the index database (e.g., update every night), and (iii) combination of (i) and (ii).

C. Search Engine

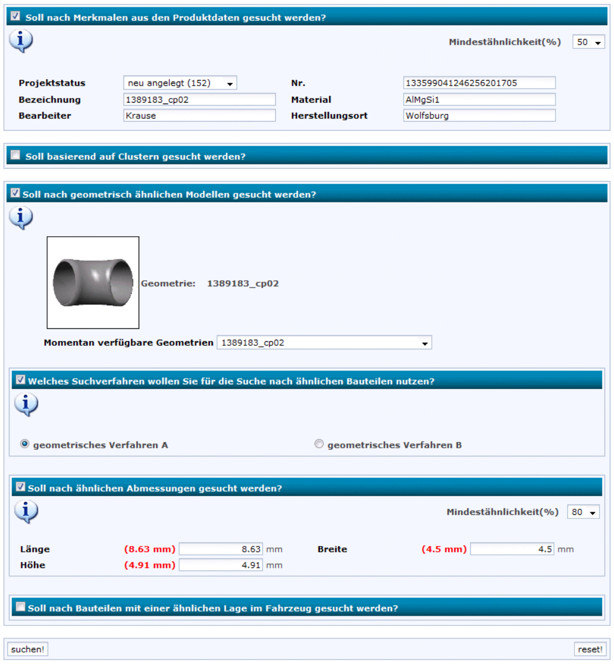
The search engine of integrated CBES implements core functionality to compare and retrieve stored enterprise process entities from the index database. Hence, an enterprise member specifies a request by defining necessary (alpha-) numerical attributes. In addition to these attributes, the enterprise member might provide template objects like 3D models or 2D sketches to describe the desired enterprise process entity (e.g., product). In short, the enterprise member submits the aforementioned information to the search engine for retrieving most similar enterprise process entities.

For retrieving most similar enterprise process entities, different similarity and distance metrics for various data types and content-based descriptors have been developed. Levenshtein distance and Smith-Waterman similarity measures are used for comparing alphanumerical data [28]. Minkowski distance or its special form i.e., Euclidean distance are used to evaluate the distance between numerical data. In addition, Hamming distance can be used to estimate the distance between two input values of equal length for numerical, binary, or string. To assess structural similarities (e.g., BOM), a graph-based distance measure has been applied [29]. To retrieve multimedia items like 3D shapes or 2D drawings, special techniques are employed. For instance, to measure shape similarity various techniques are available which can be classified into feature-based, graph-based and geometry-/image-based [18]. For content-based image retrieval, techniques from the MPEG-7 standard are utilized to detect color; texture and 2D shape similarities [30].

Enterprise process entities in the index database are added to the search result list if their calculated similarity or distance fulfills a predefined similarity threshold. The threshold value is defined by the enterprise member. This threshold value specifies the accuracy of the retrieval process. The search result list is ordered by the similarity values and returned to the enterprise member. To improve the performance of the retrieval process, the search process starts with simple attributes and descriptors, and proceeds with complex ones. For example, a user searches for a product by providing product name, 3D model and product BOM. Initially, product name and 3D model are utilized to determine a relevant subset of the data available in index database. Next, the subset of the data is revised with the structure based (e.g., BOM) search.

D. Graphical User Interfaces

The graphical user interfaces (GUI) functionality is twofold: (i) provide necessary forms for specification and submission of search requests (s. Fig. 4) and (ii) visualization of search results (s. Fig. 5). The aforementioned functionality of the GUI has to consider the roles and privileges of the enterprise members. Overall, this ensures the adherence of enterprise's IT policies and the filtering of the information and knowledge according to the enterprise members' needs.



☑ Soll nach Merkmalen aus den Produktdaten gesucht werden?

Mindestähnlichkeit(%) 50

Projektstatus	neu angelegt (152)	Nr.	133599041246256201705
Bezeichnung	1389183_cp02	Material	AlMgSi1
Bearbeiter	Krause	Herstellungsort	Wolfsburg

☐ Soll basierend auf Clustern gesucht werden?

☑ Soll nach geometrisch ähnlichen Modellen gesucht werden?

Geometrie: 1389183_cp02

Momentan verfügbare Geometrien 1389183_cp02

☑ Welches Suchverfahren wollen Sie für die Suche nach ähnlichen Bauteilen nutzen?

● geometrisches Verfahren A ○ geometrisches Verfahren B

☑ Soll nach ähnlichen Abmessungen gesucht werden?

Mindestähnlichkeit(%) 80

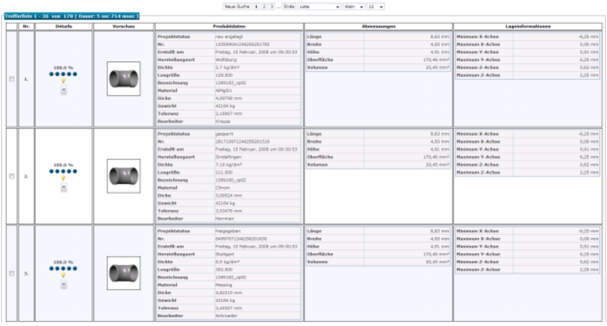
Länge	(8.63 mm)	8.63 mm	Breite	(4.5 mm)	4.5 mm
Höhe	(4.91 mm)	4.91 mm			

☐ Soll nach Bauteilen mit einer ähnlichen Lage im Fahrzeug gesucht werden?

Figure 4. Illustration of an search interface for sales personnel searching for products using metadata and 3D model

V. Case Study – Integrated CBES for Cost Estimation in Automotive Supplier Industry

The proposed architecture has been validated in industrial scenarios with emphasis on cost estimation processes of automotive suppliers [31][32]. Usually, cost estimation process starts with a request by the customer for information from the supplier about cost, manufacturing techniques and delivery date of the requested component as illustrated in Fig. 6. At supplier's side, experts have to check the component feasibility and determine the requested information (e.g., production cost). Hence, knowledge about



Trefferliste 1 - 36 von 170 (Dauer: 5 sec 714 msec)

	Nr.	Details	Vorschau	Produktdaten	Abmessungen	Lageinformationen
☐	1.	100.0 %		Projektstatus: neu angelegt Nr.: 133599041246256201705 Erstellt am: Freitag, 15 Februar, 2008 um 09:30:53 Herstellungsort: Wolfsburg Dichte: 2.7 kg/dm³ Losgröße: 129.000 Bezeichnung: 1389183_cp02 Material: AlMgSi1 Dicke: 4,69798 mm Gewicht: 42104 kg Toleranz: 2,10867 mm Bearbeiter: Krause	Länge: 8,63 mm Breite: 4,50 mm Höhe: 4,91 mm Oberfläche: 170,46 mm² Volumen: 20,45 mm³	Minimum X-Achse: -6,25 mm Maximum X-Achse: 0,00 mm Minimum Y-Achse: 0,91 mm Maximum Y-Achse: 6,25 mm Minimum Z-Achse: 0,62 mm Maximum Z-Achse: 2,25 mm
☐	2.	100.0 %		Projektstatus: gesperrt Nr.: 281710971246256201518 Erstellt am: Freitag, 15 Februar, 2008 um 09:30:53 Herstellungsort: Sindelfingen Dichte: 7.19 kg/dm³ Losgröße: 211.800 Bezeichnung: 1389180_cp02 Material: Chrom Dicke: 3,09524 mm Gewicht: 42184 kg Toleranz: 3,53478 mm Bearbeiter: Herrman	Länge: 8,63 mm Breite: 4,50 mm Höhe: 4,91 mm Oberfläche: 170,46 mm² Volumen: 20,45 mm³	Minimum X-Achse: -6,25 mm Maximum X-Achse: 0,00 mm Minimum Y-Achse: 0,91 mm Maximum Y-Achse: 6,25 mm Minimum Z-Achse: 0,62 mm Maximum Z-Achse: 2,25 mm
☐	3.	100.0 %		Projektstatus: Freigegeben Nr.: 849970771246256201638 Erstellt am: Freitag, 15 Februar, 2008 um 09:30:53 Herstellungsort: Stuttgart Dichte: 8.5 kg/dm³ Losgröße: 382.800 Bezeichnung: 1389182_cp02 Material: Messing Dicke: 4,62210 mm Gewicht: 42184 kg Toleranz: 2,44327 mm Bearbeiter: Schroeder	Länge: 8,63 mm Breite: 4,50 mm Höhe: 4,91 mm Oberfläche: 170,46 mm² Volumen: 20,45 mm³	Minimum X-Achse: -6,25 mm Maximum X-Achse: 0,00 mm Minimum Y-Achse: 0,91 mm Maximum Y-Achse: 6,25 mm Minimum Z-Achse: 0,62 mm Maximum Z-Achse: 2,25 mm

Figure 5. Result page for sales personnel showing similar products corresponding to search request submitted as shown in Fig. 4

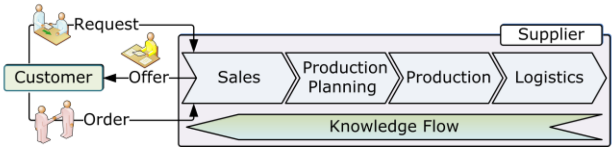


Figure 6. Typical customer supplier interaction in automotive industry

manufacturing techniques, materials and logistics are essential to perform the aforesaid tasks. Traditionally, experts from different departments support to determine the information. The data generated and used during the cost estimation process is not only textual but also multimedia-based. For example, a customer request often contains 2D drawings and / or 3D models of the requested component to be manufactured.

In the aforementioned scenario, it was determined that the existing search capabilities provided by the enterprises' IT-systems are insufficient to fulfill the requirements of an expert to identify information relevant to the cost estimation process (e.g., previously stored orders, process plans). To overcome these shortcomings, integrated CBES was developed which incorporates data from various IT-systems for previously executed offers, orders, 2D drawings and 3D models. Textual and numerical information along with multimedia information extracted from 3D models and 2D drawings were stored in the index database of the integrated CBES. This provides access to a search space containing data from various enterprise's IT-systems. To support different phases of the cost estimation process, several search interfaces have been designed. According to an enterprise member's roles and privileges, the interfaces grant access to different areas of the search space.

VI. Conclusions

In this contribution, architecture has been elaborated to provide access to enterprise's data sources through an integrated content-based enterprise search (CBES). The envisaged architecture bridges the gap between standard ES technology which is mostly focused on textual data and dedicated retrieval systems for multimedia data. The architecture consists of four components: (i) data aggregation contains techniques for indexing and descriptor creation, (ii) index database stores previously predefined descriptors, (iii) search engine implements numerous retrieval techniques, and (iv) GUI exists to submit search requests and view result lists. The developed architecture has been used in several enterprises. These enterprises are operating in different domains like automotive supplier industries and industrial chain manufacturers. The integrated CBES supports domain experts during design and execution of enterprise processes.

The effort to create an index database, implement required interfaces and integrate the required techniques for integrated CBES induces efforts (e.g., time, cost). However, the benefits of a quick and reliable access to information and associated knowledge contained in enterprises data are essential to retain an enterprise's competitive advantages. Overall, the integrated CBES improves the design and execution of enterprise processes. The throughput of enterprise processes can be increased significantly and at the same time, the quality of the enterprise processes' output can be strengthened.

Acknowledgment

The authors would like to thank their industrial partners LINDE + WIEMANN GmbH KG, Germany, Fischer & Kaufmann GmbH & Co. KG, Germany and KettenWulf Betriebs GmbH, Germany for the opportunity to implement and evaluate the elaborated architecture.

References

[1] M. Alaavi and D. Leidner, "Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues," MIS Quarterly, vol. 25, no. 1, 2005, pp. 107-136.

[2] J. Stark, Product Lifecycle Management: 21st Century Paradigm for Product Realisation, Springer, London, 2005.

[3] D. Hawking, "Challenges in enterprise search," Proc of the 15th Australasian Database Conference, vol. 27, 2004, pp. 15-24.

[4] D. V. Vranic, 3D Model Retrieval, Dissertation, University of Leipzig, 2004.

[5] J. M. Martinez, R. Koenen, and F. Pereira, "MPEG-7: the generic multimedia content description standard, part 1," IEEE Multimedia, vol. 9, no. 2, 2002, pp. 78-87.

[6] U. Crenze, S. Köhler, K. Hermsdorf, G. Brand, and S. Kluge, "Semantic descriptions in an enterprise search solution," in G. Antoniou et al. Eds., Reasoning Web 2007, LNCS 4636, Springer, pp. 334-337, 2007.

[7] M. Buttler, The Bussiness Value of Enterprise Search, Martin Butler Research, July 2009.

[8] Y. Fu, R. Xiang, M. Zhang, Y. Liu, and S. Ma, "A PDD-based searching approach for expert finding in intranet information management," in H.T. Ng et al. (eds.), AIRS 2006, LNCS 4182, Springer, 2006, pp. 43-53.

[9] P. Dmitriev, P. Serdyukov, and S. Chernov, "Enterprise and desktop search," Proc. of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 2010.

[10] L. Owens, The Forrester Wave: Enterprise Search, Q2 2008. Report, Forrester, 2008.

[11] E. K. Lee and W. Noah, "An enterprise intelligence system integrating WWW and intranet resource," Proc. of the Ninth International Workshop on Research Issues on Data Engineering: Information Technology for Virtual Enterprises (RIDE '99), March 1999, pp. 28-35.

[12] D. R. Throop, "Enterprise Search Tasks in IVHM Practice," Proc. of IEEE Conference on Aerospace, 2006.

[13] R. Winnemöller, "Semantic Enterprise Search (but no Web 2.0)," Proc. of 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Mexico City, Mexico, May 2009.

[14] Y. Yao, L. Lin, and J. Dong, "Research on ontology-based multi-source engineering information retrieval in integrated environment of enterprise," Proc. of International Conference on Interoperability for Enterprise Software and Applications China (IESA '09), April 2009, pp. 277-282.

[15] S. R. Kruk, S. Grzonkowski, A. Gzella, and M. Cygan, "DigiMe - Ubiquitous Search and Browsing for Digital Libraries," Proc. of 7th International Conference on Mobile Data Management (MDM'06), 2006.

[16] J. Bahrs, B. Meuthrath, and K. Peters, "Information Retrieval Services for heterogeneous Information Spaces," Proc. of I-KNOW '08 and I-MEDIA '08 Graz, Austria, 2008.

[17] Y. Alemu, J. Koh, M. Ikram, and D. Kim, "Image retrieval in multimedia databases: A survey," Proc. of 5th International

Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.

[18] J. W. H. Tangelder, and R. C. Veltkamp, “A survey of content based 3D shape retrieval methods,” Multimedia Tools and Applications, vol. 39, 2008, pp. 441-471.

[19] J. Pu, Y. Kalyanaraman, S. Jayanti, K. Ramani, and Z. Pizlo, “Navigation and discovery of 3D models in a CAD Repository,” Computer Graphics and Applications, IEEE , vol. 27, no. 4, July-Aug. 2007, pp. 38-47.

[20] A. Ruifrok, A. Scheenstra, and R. C. Veltkamp, “A survey of 3D face recognition methods,” Audio- and Video-based Biometric Person Authentication (AVBPA 2005), LNCS 3546, 2005, pp. 891-899.

[21] P. Geetha and V. Narayanan, “A survey of content-based video retrieval,” J. of Computer Science, vol. 4, no. 6, 2008, pp. 474-486.

[22] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, ”Content-Based Music Information Retrieval: Current Directions and Future Challenges,” Proc. of the IEEE Advances in Multimedia Information Retrieval, vol. 96, no. 4, 2008, pp. 668-696.

[23] N. Gronau, “Modelling knowledge intensive engineering processes with the knowledge modeler description language KMDL,” In: F. Weber, S. Kulwant, K. D. Thoben, (eds.) Enterprise Engineering in the Networked Economy, University of Nottingham, 2003, pp. 195-202.

[24] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, “The R*-tree: An efficient and robust access method for points and rectangles,” SIGMOD Conference, 1990, pp. 322-331.

[25] P. Ciaccia, M. Patella, and P. Zezula, “M-tree: An Efficient access method for similarity search in metric spaces,” Proc. of 23rd International Conference on Very Large Data Bases, 1997, pp. 426-435.

[26] C. Böhm, S. Berchtold, and D. A. Keim, “Searching in high-dimensional spaces - index structures for improving the performance of multimedia databases,” ACM Computing Surveys, vol. 33, no. 3, 2001, pp. 322-373.

[27] U. Müller, T. Barth, and B. Seeger, “Accelerating the Retrieval of 3D Shapes in Geometrical Similarity Search using M-tree-based Indexing,” Proc. of Workshop of Uncertainty, Knowledge Discovery and Similarity in Case-base Reasoning (UKDS’09), 2009.

[28] D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, 1997.

[29] C. J. Romanowski and R. Nagi, “On comparing bills of materials: A similarity/distance measure for unordered trees,” IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans vol 35, no. 2, pp. 249-260, 2005.

[30] B.S. Manjunath, P. Salembier, and T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley & Sons, March 2003.

[31] C. Lütke Entrup, T. Barth, and W. Schäfer, “Towards a process model for identifying knowledge-related structures in product data,” Proc. of the 6th Int. Conf. on Practical Aspects of Knowledge Management, LNCAI, vol. 4333, 2006, pp. 189-200.

[32] S. Karadgi, U. Müller, D. Metz, W. Schäfer, and M. Grauer, “Cost estimation of automotive sheet metal components using knowledge-based engineering and case-based reasoning,” IEEE Int. Conf. on Ind. Eng. and Eng. Management, Hong Kong, 2009, pp. 1518-1522.

The Strategic Alignment of Supply Chain and IT Resources

Philippe Marchildon and Pierre Hadaya
Department of Management and Technology
ESG-UQAM
Montréal, Canada
e-mail: marchildon.philippe@courrier.uqam.ca
hadaya.pierre@uqam.ca

***Abstract*—In reaction to recent findings, which suggest that when acting alone supply chain and IT resources cannot yield a competitive advantage to organizations, the present study, posits that it's by combining these resources together that organizations will gain a competitive advantage from them. Rooted in the resource-based-view and the relational view of the firm, this research presents a conceptual model that will permit the uncovering of the dominant configurations of supply chain and IT resources alignment. This study also presents three expected alignment configurations (i.e., cost driven organizations, value driven organizations and innovation driven organizations) which represent the primary form of alignment between these resources and their respective impact on organizational performance. Using the gestalt approach, we plan to verify our research model by collecting data from at least 200 Canadian prime manufacturers. Findings tied to this initiative will provide important contributions to both research and practice.**

Keywords- information management; supply chain resources; IT resources; alignment

I. Introduction

Organizations confronted with always increasing consumer demands [1] are now forced to realize tasks difficult to accomplish alone. Firms are thus relying more and more on their partners to successfully fulfill market demands [2]. This new dynamic is modifying the links bounding a firm to its partners [3] and brings the firm to outsource activities in which it is less competent [4]. Consequently, organizations are still relying on their own resources but also, and increasingly, on those accessed via their business relationships [5]. In turn, an important part of the competition among firms now occurs through their supply chain and not between individual organizations as it used to [6]. Another key factor of today's reality is that information technologies (IT) can provide competitive advantages to organizations [7]. This is particularly true in the context of a supply chain where information systems allow information to flow quickly and transparently across multiple interorganizational boundaries making it visible to all supply chain partners and in turn improving the performance of business relationships [7].

However, recent findings from IS studies question the value of these systems, arguing that they have become easily imitable necessities [8, 9]. As stated by the "resources based view of the firm (RBV)" [10] and the "relational view of the firm (RV)" [2], a firm's resources, whether housed by the firm or embedded in its relationship with its partners, cannot provide a competitive advantage if they are commonly available [11]. According to these theories, a firm will be able to obtain a sustainable competitive advantage over their competitors by combining its resources in unique and inimitable ways [12]. Furthermore, contingency theory stipulates that firm resources are ideally combined when the formers are aligned along their respective needs, demands, goals, objectives and structures. Such considerations bring managers to judiciously choose two types of key resources when establishing their business strategy: (1) those pertaining to their supply chain relationships [13] and (2) the IT resources supporting these relationships [14].

However, despite calls from several authors for firms to choose coordination mechanisms that best fit their supply chain relationship and capitalize on IT to improve their performance [15], little information is available in the literature for firms to address such alignment concerns. Indeed, even though the literature on information technology alignment at the organizational level is abundant [16], little is known at the interorganizational level. In fact, to the best of our knowledge, only four studies [17, 18, 19, 20] have attempted to investigate alignment at the interorganizational level and none of them have specifically focus on the alignment between supply chain and IT resources. It is crucial to study such alignment practices as information technologies are becoming more and more ubiquitous and easy to access due to the emergence of common communication protocol and web-based approaches. Therefore, information technologies are unlikely to provide competitive advantages by themselves and it is only trough their combination with other resources that organizations will derive benefits from them [8, 9]. Also, since competition in today's economy is now at the supply chain level it is essential to extend our knowledge on IT resource combination to the network level.

To partially address this gap in the literature, the following research aims to answer the following research question: According to the type of resources exchanged via their supply chain relationships, which information technologies will make it possible for companies to improve their performance and gain a competitive advantage? To do so, this study will develop a typology based on the dominant supply chain and IT resource alignment configurations.

The paper proceeds as follow: First, we present the literature on organizational resources (i.e., the RBV and its complement the RV). Second, we expose the underlying assumptions of the research after which the conceptual model and related propositions are presented. Third, our intended methodological actions are described. Finally, our anticipated contributions both theoretical and practical are discussed.

II. Theoretical Development

A. Two Theories on Organizational Resources: the Resource Based View and the Relational View of the firm

Two theoretical perspectives that convincingly address the complementarities between supply chain and IT resources are the RBV and its complement the RV. According to the RBV, firms can be conceptualized as "resource bundles" [21] which may earn greater profits then their rival if they are able to identify and acquire resources that are crucial in the development of demanded product or services [2]. Barney[10], in his articulation of the RBV theory, formulates two fundamental assumptions: (1) that resources and capabilities are heterogeneously distributed among firms and (2) and are imperfectly mobile. Taken together, these assumptions allow for differences in firm resource endowments to both exist and persist over time, thereby allowing for a resource-based competitive advantage [10]. The RBV also posits that it is not all resources that can provide a sustainable competitive advantage. In order to play such a role resources must meet five criteria or possess five key characteristics [10, 22]. As explained by [22, p. 1087] "First, the resource must be valuable in that it improves firm efficiency and/or effectiveness. Second, the resource must be rare so that by exercising control over it, the firm can exploit it to the disadvantage of its competitors. Third, the resource must be imperfectly imitable to prevent competitors from being able to easily develop the resource in-house. Fourth, the resource must be imperfectly mobile to discourage the ex-post competition for the resource that would offset the advantages of maintaining control of the resource. Fifth and last, the resource must not be substitutable; otherwise, competitors would be able to identify different, but strategically equivalent, resources to be used for the same purpose"

In addition, it has been argued that even if a resource does not meet the RBV criteria when acting alone, organizations can still achieve sustainable competitive advantages by combining this resource to others [12]. To do so, organizations must combine resources in a way that is valuable to the firm, scarce, difficult to imitate and not substitutable [12]. Such combination has for effect to protect combined resources from competitive imitation by path dependencies, embeddedness, casual ambiguity about the source of competitive advantage, and time diseconomies of imitation [10] and thus making them a potent source of sustainable competitive advantage. Therefore, combining resources becomes particularly important when organizations rely on resources, which have relatively low barriers of imitation and acquisition [8].

Extending the RBV, the RV posits that an organization's critical resources not only include those housed within its limits but also those imbedded it their business relationships [23]. Similarly to housed resources, those exceeding firm boundaries must also meet resource-based-view's criteria to provide a sustained competitive advantage. Furthermore, firms can also choose to combine these resources if they do not meet these criteria alone [2]. Consequently, organizations can combine resources not only at the intra-firm level but also at the interorganizational level. More precisely, [2] argue that by developing partnerships ranging from transactional to collaborative partnering organizations can combine their respective resources to create synergies between them which in turn increase barriers to imitation and allow firms to benefit from sustained competitive advantages. Synergy creation mechanisms include: (1) information/knowledge exchange, (2) presence of complementary strategic and combination of organizational resources or capability, (3) investments in relation-specific asset, and (4) effective relational governance [2]. These mechanisms preserve sustained competitive advantages derived from these combined resources by increasing causal ambiguity, time compression diseconomies, interorganizational asset interconnectedness, and by the scarcity of potential partner, resource indivisibility, and institutional environments [24].

III. Conceptual Framework

A. Underlying Assumption: The Need to Bundle Together Supply Chain and IT Resources

Despite the acknowledged importance of supply chain and IT resources, either housed or embedded in business relationships, for organizational success [25], recent studies show that each of these types of resources alone cannot yield sustained competitive advantages to organizations [22]. First, the supply chain literature indicates that supply chain relationships, and in turn the various resources associated with them, tend to provide only temporary competitive advantages to organizations since they are becoming more and more easily imitable due to technological evolutions which diminish transaction cost and encourage organizations to establish relationships with their external partners [22]. As such, organizations not only need to identify the critical resources embedded in its business relationship but also how to protect them from the mimetic behaviour of their competitor [22]. Insights from recent studies suggest that combining supply chain resources with other resources could be an adequate mean to alleviate their imitability [7, 24]. Indeed, findings from [24], in accordance to the premises of the RBV and RV, indicate that firms which invest in complementary resources to support supply chain resources can increase imitation barriers associated with them and in turn yield competitive advantages from what were at the start easy to imitate resources. Furthermore, findings from [7] also show that, by being combined with other organizational resources such as supply chain management information systems, the performance impact of relationship resources can be increased.

Second, research findings from the IS literature also indicate that IT resources may not meet the RBV criteria when acting alone [26]. More precisely, these resources, as demonstrated by [8], present relatively low barriers to imitation and acquisition by other firms making IT-based advantages to diminish rather quickly over time [26]. Consequently, having recognised the limits of IT resources, authors from the IS field investigated various ways by which organizations could derive a sustainable competitive advantage from IT resources [26]. Following this endeavour, some authors have argued that one approach to palliate to IT resources shortcomings consist of judiciously combining them with other organizational resources [11] such as supply chain resources [7]. Indeed, by combining their IT resources with other organizational resources, firms could be able to insulate these resources from competitive imitation. Key findings from [11] corroborate this claim by indicating that the value of IT can be augmented only when it is embedded in an organization through resource complementarities and co-specialization.

Taken together, these streams of research argue that each type of resources would gain from an appropriate combination with other resources. Furthermore, organizations should seek to combine both types of resources together in order to minimize their respective shortcomings trough complementarities [26]. Indeed, the literature in supply chain management recognizes that information technologies represent a critical driver of supply chain success [27] while the IS literature recognizes that the full potential of IT resources can only be obtained through their adequate combination with other organization resources such as supply chain resources [7, 24, 26]. Hence, the main assumption of this research is that, to achieve a sustainable competitive advantage, organizations should align their supply chain and IT resources.

B. *Research model:Uncovering the Dominant Configurations of Supply Chain and IT Resources Alignment*

In accordance with our underlying assumption, Fig 1 exposes our research model, which will be used to uncover the dominant configurations of supply chain and IT resources alignment. The model comprises three facets (supply chain resources, IT resources and organizational performance.), which are detailed in the next sub-sections.

1) *Supply chain resources:*

From a RBV perspective, a supply chain relationship or supply chain linkage – defined as an “explicit and/or implicit connections that a firm creates with critical entities of its supply chain in order to manage the flow and/or quality of inputs from suppliers into the firm and of outputs from the firm to customers [22, p. 1084]” – can be seen as a resource per see or as a capability that allows a firm to acquire resources which in turn can yield benefits [22]. Although both viewpoints recognize the importance of supply chain linkage and are congruent with the RBV, they differ considerably on how and why they may provide sustainable competitive advantages to firms [22]. The former presumes that simply having a critical link with one supply chain partner guaranties some sort of abnormal rent while the later posits that even if an organization is linked to its partner, the firm still needs to exploit this relationship by acquiring or sharing resources with its partner to obtain a sustained competitive advantage.

As mentioned previously, firms are changing the nature of their relationship with their supplier; customers and other external partners forming new interorganizational coalitions, such as virtual enterprises and integrated supply chains [3]. Such changes stem from organizations’ desire to move away from arms-length relationships to more collaborative partnerships [3] and harness benefits from closer and stronger partnerships [28]. In order words, organizations recognize that the simple fact of establishing a partnership (arm’s length relationship) is not a guaranty for success, and that relationship should be viewed as a mean to efficiently manage forward flow of material and backward flow of information. The underlying aim of this observed organizational behaviour to encourage supply chain linkage forces us to consider supply chain relationships as capabilities, which allow organizations to acquire or share resources, and not as a resource per see.

Insights from case studies indicate that organizations, which establish supply chains relationships, do so in order to efficiently manage or acquire three different types of resources (1) materials, financial and information [7]. However, of these three resources, information and its effective management across supply chain partners is the one that exerts the greater impact on firm performance [7]. As such, the present research only focuses on this particular supply chain resource. Information sharing between supply chain partners has been examined by scholars from diverse background including, among others, information systems, operation management and marketing [24]. Following an extensive literature review on the subject, [29] concluded that information flows or information was a multi-dimensional resource encompassing three different sub-set of information: (1) operational, (2) tactical and (3) strategic each affecting organizational benefits differently and positively when shared efficiently and effectively [24]. Drawing from the work of [29], the present research adopts a three level classification framework for information and differentiates between operational, tactical and strategic information levels. More precisely, operational information refers to information tied to the production of product and services, such as information about resources conditions and plans such as inventory/capacity plans and production schedules [24]. When shared efficiently, this information allows partnering organizations to optimize input resources globally by streamlining buffers and synchronizing resource allocations [24]. As such, organizations sharing this type of information can achieve operational economies-of scale and reduce inventory and ordering cost [30].

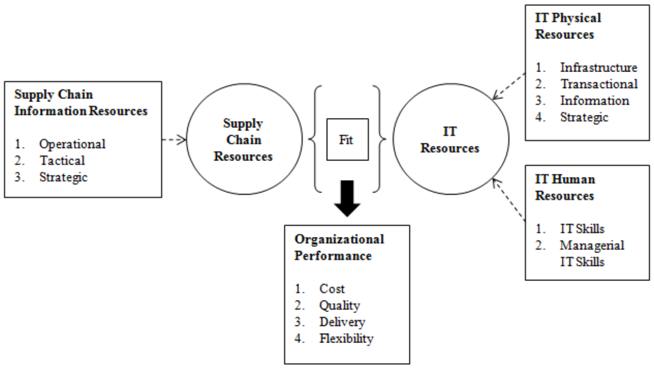


Figure 1. Poposed research model

Tactical information relates to financial metrics on margin structures and costs [24]. When shared adequately, this information enables parties to collaborate on ways to improve economic outcomes and to leverage both parties' resources [24]. As such, organizations sharing this type of information are usually able to improve their response to customer demands trough adequate delivery practices (i.e., continuous replenishment and quick response systems), making the flow of material in the supply chain to be "pull" by consumer demands rather than "pushed" by producers [30]. Finally, strategic information is defined as information that affects a firm competitive positioning and planned actions in the market [24]. When shared efficiently, this information allows business partners to obtain or increase their benefits by coordinating sales and marketing activities with operational requirements [24]. Therefore, allowing organizations sharing this type of information to move into new markets or develop new product [7].

2) IT Resources

Despite the fact that the RBV provides a helpful theoretical lens from which to assess the role of IT resources and their business value [31], the existing literature in the IS field is rather ambiguous on their definition and conceptualization [32]. For example, many different classification schemes have been proposed [9, 11, 31, 33, 34, 35]. From these proposed classification, two general conclusions can be drawn. First, IT resources are not monolithic and thus encompass different dimensions. Second each framework usually distinguishes between two types of complementary IT resources: physical IT resources and human IT resources, which are also consistent with Grant's classification scheme for resources [31, 33]. Both types of IT resources, physical and human, are considered essential and each complementarily enhances the success of a firm [32].

Physical resources usually refer to infrastructure and related deployment resources [31]. They are considered to be multi-dimensional and can include different types of IT investment: innovative vs. non-innovative, strategic vs. nonstrategic, and internally focused vs. externally focused investments [36] each reflecting a firm's strategy and affecting its performance accordingly [32]. The classification that most convincingly addresses this multi-faced role of IT physical resources is the one proposed by [32] which distinguishes between four different types of IT physical resources or IT investments: infrastructure, transactional, information, and strategic investments.

Infrastructure investments relate to shared IT services such as servers, networks, laptops, shared customer databases, help desk and application development used by multiple IT application. This type of IT resources provides the groundwork for present business initiatives as well as a flexible base for future business initiatives [32]. They require high up-front costs, which are in turn outbalanced by long-term performance improvements [32]. Transactional investments refer to investment made with the aim of automating repetitive business transactions and processes such as order processing, point of sale processing, bank cash withdrawal, billing statement production and other repetitive transactions [32]. As such, this type of investment is likely to cut organizational costs and/or increase the volume of business a firm can conduct per unit cost [32]. Information investment provides information to firm's managers communicating internally and externally with their supply chain partners. These investments can take the form of decision support systems that enable more effective decision making by allowing sale analysis and data mining. These types of investment influence a firm performance on the following indicator: control, reliability, delivery and adaptability of firms [32]. Strategic investment refers to IT resources which help repositioning the firm into the marketplace whether by supporting a firm's entry into new markets or by enabling the development of new products, services or business processes [32]. Consequently, these investments are likely to increase the flexibility of an organization in regards to customer demands.

Human IT resources, similar to physical resources, are multi-facet and are recognized to include technical and managerial IT skills [35]. Technical skills refer "to the know-how needed to build IT applications using the available technology and to operate them to make products or provide services" [9, p. 498]. These skills allow employees to be more productive which in turn decrease costs and improve other operational performance indicators [9, 35]. Furthermore, technical skills also enable firms to efficiently manage the technical risk associated with infrastructure investment [9], which in turn also diminish organizational costs. On the other hand, managerial skills refer to the "management's ability to conceive of, develop, and exploit IT applications to support and enhance other business functions" [9, p. 498]. These skills, compared to technical skills, relate more to employee's communication and analysis abilities. More precisely, they allow employees: (1) to better understand and appreciate the business needs of their counterparts, both internal and external, (2) to work with them in developing appropriate IT applications, (3) to coordinate IT activities in ways to support each other, and (4) to anticipate the future IT needs for all partners [9]. Consequently, managerial IT skills help organizations to reap the full potential of IT by increasing the adaptability of its employee, which in turn improve the flexibility of the organization and its level of customer service [9, 33].

3) Organizational performance

By establishing relationships with their trading partners, organizations aim to successfully answer final customers' demands [37]. Such demands from consumers are usually formulated along four evaluation criteria: price, quality, delivery and availability [1]. Accordingly, supply chain performance has usually been assessed along the corresponding criteria of cost, quality, delivery and flexibility [37, 38]. Cost relates to production cost, productivity, capacity utilization and inventory reduction while delivery criteria include: on-time delivery, short-time delivery, production lifecycle, lead-time and delivery on due date [39]. On the other hand, customers usually assess quality along eight dimensions: performance, features, reliability, conformance, durability, serviceability, aesthetic and perceived quality, from which the last two are inherently complex and the most difficult to measure [39]. Flexibility refers to a supply chain's agility, adaptability, and responsiveness to the needs of its users [40] and can be assessed along three dimensions: product mix, volume change over and modification [ward].

C. Research Propositions :Primary Forms of Supply Chain and IT Recources Alignment

In the context of this conceptual paper, we present, in this section, three expected alignment configurations that represent the simplest form of successful alignment between these resources to improve the various dimensions of organizational performance.

1) Configuration 1: Cost driven organizations

The first basic configuration proposed is anchored around operational information and its efficient management through its combination with key IT resources. As mentioned previously, operational information refers to information tied to the production of product and services and includes information about resources conditions and plans [24]. This type of information is, by nature, rather repetitive and requires limited interpretation, thereby making transactional investments a perfect match since they allow organization to automate repetitive business transaction [32]. However, other IT resources will also be needed as transactional investments also require infrastructure investments and technical IT skills to be efficient. Indeed, infrastructure investments provide the backbone from which every other IT investment is anchored [32] while technical IT skills allow users to efficiently use these transactional investments [9]. The combination of these four distinct resources is likely to yield a sustained competitive advantage to organization based on costs differentiation. Indeed, (1) sharing operational information allows economies-of scale, while reducing inventory and ordering costs [30], (2) transactional investments automate the sharing of operational information driving costs further down, (3) technical IT skills increase productivity which in turn also decrease costs [9, 35], and (4) infrastructure investments, when combined with technical IT skills which diminish implement costs, are also tied to cost reduction [32].

P1: When an organization mainly exchanges operational information with its supply chain partner and the operational information is combined with IT infrastructure investments, transactional investments and technical skills, the cost performance of the organization will improve.

2) Configuration 2: Value driven organizations

The second configuration presented here relates to the effective management of tactical information. Tactical information focuses on financial metrics, margin structures and costs [24]. This type of information is meant to help partners to collaborate and leverage their respective resources [24]. In turn, IT information investments are well suited to support the collaboration between partners by enabling internal and external communication between partners [24]. As such, we expect organizations to combine these two complementary resources together. Organizations relying on this combination will also need to add three other IT resources: infrastructure investments, technical skills and managerial skills. Again infrastructure investments are necessary to procure the adequate hardware required by information investments and technical skills will allow efficient use of these investments [9, 32]. Managerial IT skills are also essential since they represent communication and analysis abilities which are key when collaborating or when customizing decision support systems [9]. Taken together, tactical information, information investments, infrastructure investments, technical skills and managerial skills allow better decisions and in turn improve organization response to customer demands [30, 33] These improvements can take the form of demand driven supply chains or increase customer service [30, 33]. Thereby, the alignment of these resources should yield a sustained competitive advantage to organization based on quality and delivery differentiation.

P2: When an organization mainly exchanges tactical information with its supply chain partner and the tactical information is combined with infrastructure investment, information investment, technical IT skills and managerial IT skills, the quality and delivery performances of the organization will improve.

3) Configuration 3: Innovation driven organizations

The last configuration proposed focuses on strategic information sharing. Organizations share strategic information to position themselves in the market and coordinate related actions [24]. Such actions can take the form of new product development and entry into new markets [7]. Strategic investments also have the same objective as they are destined to support similar actions [32] making them an ideal support to strategic information. As such, we expect organizations to match these resources together. This combination will also require three complementary IT resources: infrastructure investments, technical IT skills and managerial IT skills. Infrastructure investments and technical IT skills will play the same roles as previously described in configuration two. Managerial IT skills will permit managers to easily adapt to and anticipate future IT needs [9, 33] and thus enhance the value of organizational positioning actions. These resources, by being combined together, will allow organizations to anticipate and effectively reply to changes in customer demands by facilitating new product development and entry into new markets [9, 24, 32, 33], As such, we expect organizations pertaining to this configuration to gain a sustained competitive advantage based on flexibility differentiation.

P3: When an organization mainly exchanges strategic information with its supply chain partner and the strategic information is combined with infrastructure investment, strategic investment, technical IT skills and managerial IT skills, the flexibility performance of the organization will improve.

IV. RESEARCH METHODOLOGY

A. Data Collection

We plan on validating our research model with a stratified sample of 200 critical prime manufacturer-supplier relationships, where prime manufacturers are located in Canada and active in the four following industrial sectors: (1) machinery manufacturing, (2) computer and electronic product manufacturing, (3) electrical equipment, appliance and component manufacturing and (4) transportation equipment manufacturing. Top executive responsible of the supply chain activities of each manufacturer will be the selected respondent. For each respondent, we will collect information on a single buyer-supplier relationship, but the name of the chosen supplier need not be provided.

B. Reasearch Construct and Measures

Some constructs of the conceptual model have been previously used by researchers in the field of IS or supply chain (i.e., IT infrastructure resources [32], IT transactional resources [32], IT informational resources [32], IT strategic resources [32], cost performance [39, 40], quality performance [39, 40], delivery performance [39, 40] and flexibility performance [39, 40]) while others (i.e., operational information, tactical information, strategic information, IT skills, and Managerial IT skills) will be developed using [41] paradigm for measure development.

C. Satistical Analyses

For the purpose of this the study the gestalt perspective will be employed as it allows the uncovering of typologies (configurations) which is the major aims of this research. More precisely, this study will follow [17] six steps analytical process to find the configurations of supply chain resources and IT resources alignment. One-way analysis of mean (ANOVA) will also be used to identify the best performing configurations.

V. CONCLUSION

Rooted in the RBV and the RV, this research proposes a model that will permit the uncovering of supply chain and IT resources alignment configurations. Findings tied to this initiative will provide important contributions to both research and practice.

Alignment studies have traditionally been concerned with the extent of fit rather than the form of fit associated with IT resources. The present research significantly depart from these previous research endeavor and makes a significant contribution to research by being one of the few to investigate both the level and the form of alignment between supply chain and IT resources. Furthermore, contrary to most studies in the IS field, which have empirically assessed the role of IT resources at an aggregate level, this research proposes to empirically assess a set of physical and human IT resources. This will not only extend our understanding of IT resource alignment but will also increase our knowledge tied to organizational resources by revealing the distinct nature of IT resources and their respective role and impact.

From a methodological perspective, this research makes a significant contribution to research by validating a rigorous approach to cluster analysis, which extends our knowledge on alignment assessment and validation. This research will also develop important scales necessary to the measure supply chain and IT resources thereby making another important contribution to research.

From a practical viewpoint, this study will allow organizations to better manage their resources by identifying (1) their respective strengths and weaknesses, (2) their respective impact on various performance dimensions and (3) interaction effects that can entail sustainable competitive advantages.

REFERENCES

[1] J. Griffiths, R. James and J. Kempson, “Focusing customer demand through manufacturing supply chains by the use of customer focused cells: An appraisal,” International Journal of Porduction Economics, vol. 65, 2000, pp. 111-120.

[2] J.H. Dyer and H. Singh, “The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage,” Academy of Management Review, vol. 23, 4, 1998, pp. 660-679.

[3] M. Bensaou, “Interorganizational Cooperation: The role of Information Technology an Empirical Comparison of U.S. and Japanese Supplier Relations,” Information Systems Research, vol. 8, 2, 1997, pp. 107-124.

[4] M. Sobrero and E.B. Roberts, “Strategic Management of Supplier-Manufacturers Relations in New Product Development,” Research Policy, vol. 31, 2002, pp. 159-182.

[5] D. Tapscott, D. Ticoll and A. Lowy, “Digital Capital: Harnessing the Power of Business Webs,” Boston, MA, Harvard Business School Press, 2000.

[6] M. Christopher and D. Towill, “An Integrated Model for the Design of Agile Supply Chains,” International Journal of Physical Distribution & Logistics Management, vol. 31, 2001, pp. 235-246.

[7] A. Rai, R. Patnayakuni and N. Seth "Firm Performance Impacts of Digitally Enable Supply Chain Integration Capabilities," MIS Quarterly, vol. 30, 2, 2006, pp.225-246.

[8] E.K. Clemons and M.C. Row, "Sustaining IT advantage: the role of structural differences," MIS Quarterly, vol. 15, 3, 1991, pp. 275-292.

[9] F.J. Mata, W.L. Fuerst and J.B. Barney, "Information technology and sustained competitive advantage: A resource-based analysis," MIS Quarterly, vol. 19, 4, 1995, pp. 487–505.

[10] J. Barney, "Firm Resources and Sustained Competitive Advantage," Journal of Management, vol. 17, 1991, pp. 99-120.

[11] T.C. Powell, and A. Dent-Micallef, "Information Technology as Competitive Advantage: The Role of Human, Business, and Technology Resources," Strategic Management Journal, vol. 18, 5, 1997, pp. 375-405.

[12] R. Grant, "Prospering in Dynamically-Competitive Environments: Organizational Capability as Knowledge Integration," Organization Science, vol. 7, 4, 1996, pp. 375-387.

[13] A. Agarwal, R. Shankar, and M.K. Tiwari, "Modeling the Metrics of Lean, Agile and Leagile Supply Chain: An ANP-Based Approach," European Journal of Operational Research, vol. 173, 2006, pp. 211 225.

[14] K. Kemppainen and A. Vepsäläinen, "Trends in Industrial Supply Chains and Networks," International Journal of Physical Distribution & Logistics Management, vol. 33, 8, 2003, pp. 709-719.

[15] R. Bunduchi,."Trust, power and transaction costs in B2B exchanges — A socio-economic approach," Industrial Marketing Mangement, 37, 2008, pp. 610-622.

[16] F. Bergeron, L. Raymond and S. Rivard, "Ideal Patterns of Strategic Alignment and Business Performance," Information Management, vol. 41, 8, 2004, 1003-1020.

[17] M. Bensaou and N. Venkatraman, "Configurations of Interorganizational Relationships: A comparison Between U.S. and Japanese Automakers," Management Science, vol. 41, 9, 1995, pp. 1471-1492.

[18] H-L, Chang, K. Wan and I. Chiu, "Business–IT fit in e-procurement systems: evidence from high-technology firms in China," Information Systems Journal, vol. 18, 2008, pp. 381-404.

[19] P.W. Forster and A.C. Regan, "Electronic Integration in the Air Cargo Industry: An Information Processing Model of On-Time Performance," Transportation Journal, vol. 40, 4, 2001, pp. 46-61.

[20] G. Premkumar, K. Ramamurthy. and C.S. Saunders, "Information Processing View of Organizations: An Exploratory Examination of Fit in the Context of Interorganizational Relationships," Journal of Management Information Systems, vol. 22, 1, 2005, pp. 257-294.

[21] E. Penrose, "The Growth of the Firm," Wiley, New York, 1959.

[22] M. Rungtusanatham, F. Salvador, C. Forza and T.Y. Choi, "Supply-chain linkages and operational performance: A resource-based-view perspective," International Journal of Operation and Production Management, vol. 23, 9, 2003, pp. 2084-1099.

[23] W.W. Powell, K.W. Koput and L. Smith-Doerr, "Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology," Administrative Science Quarterly, vol. 41, 1996, pp. 116-145.

[24] R. Klein and A. Rai, "Interfirm Strategic Information Flows in Logistics Supply Chain Relationship," MIS Quarterly, vol. 33, 4, 2009, pp. 735-762.

[25] V. Sambamurthy, A. Bharadwaj and V. Grover, "Shaping Agility Trought Digital Options :Reconceptualizing the Role of Information Technology in Organization," MIS Quarterly, vol. 27, 2,2003, pp. 237-263.

[26] F. Wu, S. Yeniyurt, D. Kim and S.T. Cavusgil, "The impact of information technology on supply chain capabilities and firm performance: A resource-based view," Industrial Marketing Management, vol. 35, 2006, pp. 493-504.

[27] S. Raghunathan, "Interorganizational Collaborative Forecasting and Replenishment Systems and Supply Chain Implications," Decision Sciences, vol. 30, 4, 1999, pp. 1053-1071

[28] B.A. Weitz and S.D. Jap, "Relationship Marketing and distribution channels," Journal of Academy of Marketing Science, vol. 23, 4, 1995, pp. 305-320.

[29] R. Patnayakuni, A. Rai and N. Seth, "Relational Antecedents of Information Flow Integration for Supply Chain Coordination," Journal of Management Information Systems, vol. 23, 1, 2006, pp. 13-49.

[30] A. Seidmann and A. Sundararajan, "The Effects of Task and Information Asymmetry on Business Process Redesign," International Journal of Production Economics, vol. 50, 2,1997, pp. 117-128.

[31] C. Zhang and J. Dhaliwal, "An investigation of resource-based and institutional theoretic factors in technology adoption for operations and supply chain management," International Journal of Production Economics, vol. 120, 2009, pp. 252-269.

[32] S. Aral and P. Weil, "IT Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation," Organization Science, vol. 18, 5, 2008, pp. 763–780.

[33] A.S. Bharadwaj, "A resource-based perspective on information technology capability and firm performance: An empirical investigation,". MIS Quarterly, vol. 24, 1, 2000, pp. 169-196.

[34] A.S. Bharadwaj, V. Sambamurthy and R.W. Zmud, "IT capabilities: Theoretical perspectives and empirical operationalization," In: Hirschheim, R., Newman, M., Degross, J.I., Eds, Proceedings of the 19th International Conference on Information Systems, Helsinki, Finland,1998, pp.378–385.

[35] N. Melville, K. Kraemer and V. Gurbaxani, "Review: Information technology and organizational performance: An integrative model of IT business value," MIS Quarterly, vol. 28, 2, 2004, pp. 283–322.

[36] A.S. Bharadwaj, S.G. Bharadwaj and B. Konsynski,. "Information technology effects on firm performance as measured by Tobin's ," Management Science, vol. 45, 7, 1999, pp. 1008–1024.

[37] S.K. Vickery, J. Jayaram, C. Droge and R. Calantone, "The effects of an integrative supply chain strategy on customer service and financial performance: an analysis of direct versus indirect relationships," Journal of Operations Management, vol. 21, 2003, pp. 523-539.

[38] S. Devaraj, L. Krajewski and J.C. Wei, "Impact of eBusiness Technologies on Operational Performance: The Role of Production Information Integration in the Supply Chain," Journal of Operation Management, vol. 25, 6, 2007, pp. 1199-1216.

[39] P.T. Ward, J.K. McCreery, L.P. Ritzman and D. Sharma, "Competitive priorities in operations management," Decision Sciences, vol. 29, 4,1998, pp. 1035-1046.

[40] G.T.M. Hult, D.J. Ketchen Jr, S.T. Cavusgil and R. Calantone, "Knowledge as a strategic resource in supply chains," Journal of Operation Management, vol. 24, 2006, pp. 458-475.

[41] G. A. Churchill Jr. "A paradigm for developing better measures of marketing constructs," Journal of Marketing Research, vol. 16, Feb 1979; pp. 64-74, doi: 000001; ABI/INFORM Global

Industrial Application of Ontologies

Real life examples

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

***Abstract*—The industrial application of ontologies is usually connected to a real life problem. Over the last 2 years we used ontologies to solve problems in the areas of retail article management, contract and Request for proposal (RFP) analysis, standard service catalogues and materials management. All these problems were either based on insufficient knowledge of the data and information by the data owner itself, or by semantic and constraints challenges, which could not be resolved due to the complexity and size of the data. In this paper, we will explain how ontologies have been set up and which algorithms have been used to resolve these problems. The combination of ontologies with the analysis of dependencies, text structures, outliers, patterns and similarities lead to an analysis approach and - in the very end - a tool, which on one hand is simple enough to be understood by an industrial expert and on the other hand mature enough to provide the analysis features described above. This paper is more field report than a research paper, but should give an impression that there are areas of application beside the academic world the urgently require ontology based knowledge management.**

Keywords- ontology; mapping; analysis features; service catalogues; article and material management.

I. Introduction

At eKNOW 2010 we presented a paper about a research project called OPTIKON [1]. The goal of OPTIKON is to develop an ontology based methodology that allows Small and Medium Enterprises (SME) to identify requirements from different standards and map these requirements into one combined set of requirements. This should allow a SME to satisfy a high number of standards with minimal – or at least reasonable – effort.

At the end of the presentation we promised to come back to eKNOW 2011 and explain what happened over the last 12 month with this approach, what worked and what did not.

But, since February 2010 our world changed significantly. Whenever one of our customers came up with a new data problem, our first idea always was “why not trying to solve it with an ontology?” And this is where we are today – solving data problems by thinking in “ontology terms” and trying to help our customers with this approach as much as possible.

We are no researchers – so please forgive us some simplifications. But we hope that our real life examples may give you some benefit and motivation that your work is valued, used and required every day to generate economic benefit.

In section II and III we will explain our initial and final approach. Section IV describes how the required data can be collected. Section V holds a description of the implemented analysis features, whilst section VI show how results are generated from these. In section VII we will show examples of the application in industry before we end with a conclusion in section VIII.

II. From mind to hand to machine

In the beginning of our work, we tried to discuss with our customers (or better the problem owner) that we want to solve their problems by using ontologies. The reaction was opposition. For most of our customers, ontology was an over-the-top theoretical thing, good for universities and philosophers but never fit for purpose to solve their problems.

So we changed our approach – we asked them to draw bubbles. Each bubble should represent a set of information or data they are concerned about. Then we asked them to describe dependencies and interactions between these bubbles by drawing a line. In a next step we asked them whether there is more information required to understand their problem. So they draw more bubbles and connections. Once they had completed this “drawing” we asked them to describe their perfect data world with the same means: bubbles and lines. In the very end we tried to draw lines from the real world to the perfect world, which should represent procedures to resolve the problems in the perfect world.

Bringing it all together: we helped them drawing an as-is ontology, a to-be ontology and a mapping between both.

After this exercise we had some very good sheets of paper, but still one major problem: each bubble represented millions of data fields buried somewhere in a database waiting for being touched by the beauty of an ontology.

Coming from an IT background the solution was right at hand: we needed some kind of tool that allowed us to

- Draw the ontologies
- Assign data to concepts and relations
- Perform analysis on concepts and relations
- Generate results

- Correct identified inconsistencies in data

III. From mind to hand to machine

The starting point for all work was to bring the ontology from a sheet of paper to a system.

A. *Drawing as-is and to-be*

The first required feature was to draw a very basic ontology – containing just concepts (called nodes) and relations (called connections). This feature allowed to draw an ontology of the current situation (as-is) as well as of the intended end stage (to-be).

B. *From as-is to to-be*

As-is and to-be ontologies only make sense if they can be mapped. For this reason three connectors have been created: the direct relation, the split and the combination.

A direct relation ensures that a concept of the as-is ontology can be mapped to a concept of the to-be ontology by applying some set of rules(1:1 mapping).

A split relation allows to divide a concept of the as-is ontology into more than one concept of the to-be ontology (e.g., address may be “splitted” into street, number, postcode, city).

A combination relation does the same as a split relation, only in the different direction (e.g., combines elements of an address).

Due to complexity reasons, a n:m relation was not modeled. Nevertheless it is possible for most cases by a two-step approach. First combination relations (n:1) are used, then the to-be ontology becomes an as-is ontology, on which then split relations are applied (1:m).

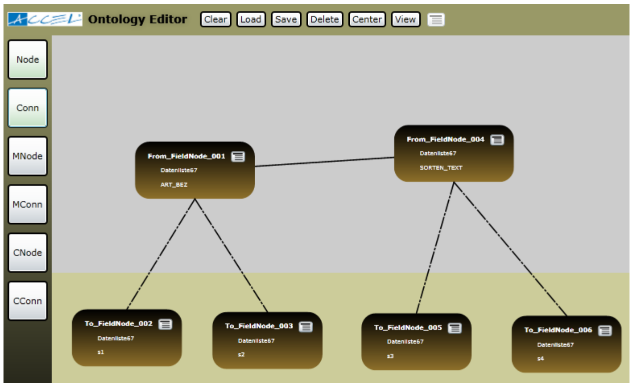


Figure 1. A very basic as-is and to-be ontology (for enlarged picture see Appendix)

IV. Getting the data

Even though the ontologies build the basis for all work, the user is more interested to see the real world picture of the individuals.

A. *Database input*

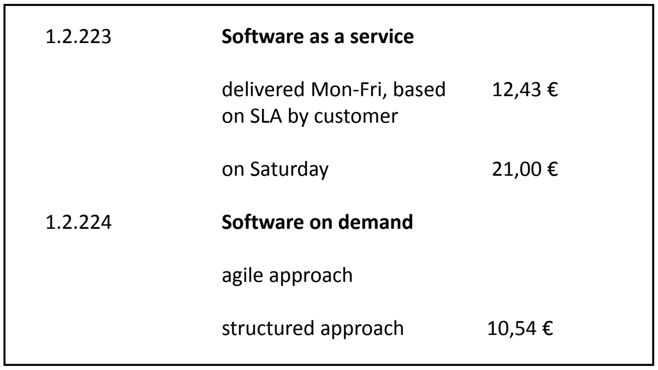
The easiest way of assigning individuals is the database input. A concept is assigned to a database column and by this all fields of this column become individuals of a concept.

B. *Text input*

A more sophisticated approach is to use text as an input. If one wants to retrieve data from text there usually are two approaches: if the text is unstructured, the user tags words or text elements that represent a concept or individual; if the text is structured, the user tries to identify the pattern of the structure and afterwards assigns concepts and individuals based on the pattern.

Both approaches have been implemented. By a tag function the user identifies words or text elements that represent a concept. Individuals are tagged in the same way and then assigned to an already existing - or newly tagged – concept.

The implementation of patterns was more challenging. If a text is structured in a pattern, it might contain some table structures, integrated pictures, etc. Therefore it was decided to analyze structured texts by graphical means. Each element of a text structure became a “box”. Based on the size and distance of the boxes, related text elements can be identified as well as maybe missing elements.



1.2.223	**Software as a service**	
	delivered Mon-Fri, based on SLA by customer	12,43 €
	on Saturday	21,00 €
1.2.224	**Software on demand**	
	agile approach	
	structured approach	10,54 €

Figure 2. Example of a structured text

Based on the approach described above, the text will be “boxed” as follows:

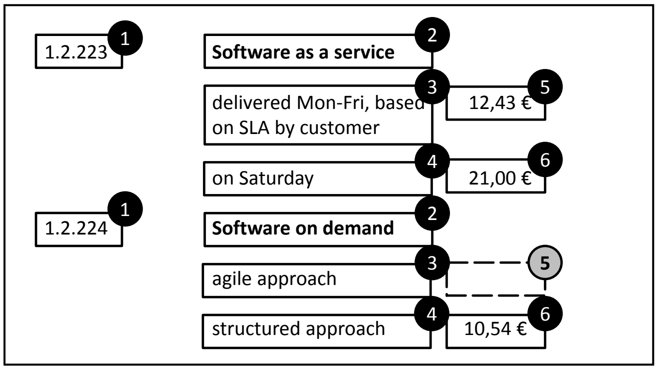


Figure 3. Boxes a structured text

C. *Supporting Tables*

Text and database input assume that individuals have to be generated from a text or database. But in some cases it might also be possible that the required set of individuals is already know and available. In this case a table is assigned to a concept, which holds only required and valid values.

As this kind of table in most cases has been used to support the validation of individuals from other tables (or concepts), it has been named “supporting table”.

V. ANALYSIS FEATURES

By now, all effort has been made to structure information (as-is / to-be ontology, relations, concepts) and assign data (individuals). But the main reason for this effort is to prepare required and intended analysis. Given our current realm of experience, the set of analysis features must at least cover the elements below.

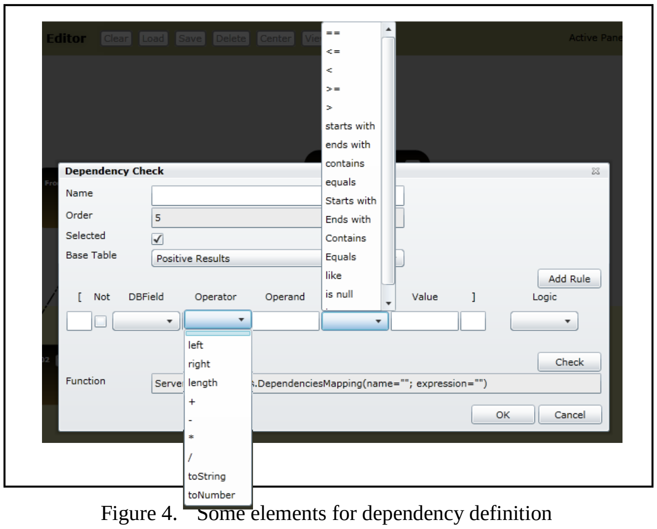
A. *Text structure analysis*

Text structure analysis looks after the structure and content of the individuals of a concept. If an activity e.g., should consist of a verb, a noun and the number of occurrences, “wash hands twice” would be a valid text structure. “Hands wash twice” will be as well invalid as “wash twice”, “twice wash hands” and “wash clean twice”.

The major means for text structure analysis are the split between as-is and to-be ontology (to split the text structure into its elements) and the supporting tables (e.g., to limit the number of occurrences).

B. *Dependencies*

Dependency analysis restricts the number of individuals per concepts. For a to-be ontology it may be required that only individuals containing a specific text or value, starting with a specific number… are allowed to be assigned to a concept. The power of this analysis is driven by the text, mathematical and pattern features applicable to a dependency.



Dependency Check

Name

Order 5

Selected ☑

Base Table Positive Results

Function Server[illegible].DependenciesMapping(name=""; expression="")

Figure 4. Some elements for dependency definition

C. *Outliers*

If a high number of individuals are assigned to a concept, it should be checked whether these individuals contain unwanted outliers, which may impact the overall result. Therefore a defined set of top or bottom elements (with regard to the mean or median, based on value or %-age) must be identifiable.

D. *Pattern*

In extension to the outlier analysis, one should be able to restrict the number of individuals based on a specific pattern. If, for example, a valid date always follows the patter 99. XXX 9999 (where 9 stands for number and X for letter), all individuals not following this pattern should be marked for rework.

E. *Duplicates*

If individuals are assigned to a concept from a database or text, usually duplicates are generated (e.g., if an address database is analyzed by the split function as described above, most cities will be assigned to the concept “city” multiple times).

To resolve this, it must be possible to identify and delete duplicates.

F. *Syntactical similarities*

A specific type of duplicates can occur if individuals had been collected or modified manually. Typos or different abbreviations may lead to different syntax for semantically identical elements (e.g., “number of elements” vs. “no of elements” vs. “nr of elements” vs. “numb of elements”).

In these cases algorithms like Levenshtein distance [2], with its extension by Ukkonen [3], the Baeza-Yates-Gonnet Algorithm [4] and others may be used.

G. *Automated correction*

Once the analysis features have identified the correct set of individuals, one has to cope with the incorrect individuals. Again here one will find 2 groups: intended incorrect and unintended incorrect.

Intended incorrect individuals are individuals that break a defined rule and for no applicable other rule this individual is valid. Unintended incorrect individuals still break a defined rule but may deliver a valid result for a parallel or corresponding rule.

If, for example, there is a rule that an address should always consist of street, number, postcode and city, “Lunen, D44536, Zum Pier, 73” has 4 unintended incorrect values as based on the rule and its split into 4 elements, none of the element delivers a valid value at its position; but by resorting the values, a valid structure can be reached.

On the other hand “Zum Pier, 73, D44536, Germany” has an intended incorrect individual as “Germany” does in no case fulfill the requirements of the rule.

To optimize the benefit of the user, for each identified incorrectness it has to be checked, whether the incorrectness can be resolved by either resorting or replacment by a valid value.

H. *Using supporting tables and majorities*

Whilst the automated correction by resorting of values (individuals / parts of individuals) can easily be realized, replacement requires a second look. To replace a value two approaches have been implemented.

The most trustful way is to use supporting tables. If a supporting table contains a value with a high syntactical similarity the original value may be replaced by this, if the calculated syntactical similarity is high enough (based on a threshold, which should be dependent on text length).

If no supporting table is available, the correct value may be identified by majority observation. If, for example, in a database two values have a high syntactical similarity, both are correct (or incorrect) at first sight, and one value occurs e.g., 10 times more often than the other value, it may be assumed that the value with the highest number of occurrences should become the correct value. This should be considered carefully, as e.g., low thresholds for syntactical similarity usually lead to inconsistent replacements.

For completeness, a third approach should be mentioned: whenever two values have a syntactical similarity that allows replacement and one of them is in the set of correct results, this value always replaces the other.

I. *Using meta notes*

Even though not mentioned explicitly yet, all analysis features are limited to 2 concepts, as the basic structure is that one relation only connects 2 concepts.

As in some cases dependency, similarity or other analysis have to be performed on more than 2 concepts at the same time (e.g., only those addresses where postcode start with “45”, city starts with “L” and street starts with “Z”) we invented a so called meta-node. A meta-node is a node that collects information from more than one relation. If the concept address from the as-is ontology is connected to street, number, postcode and city in the to-be ontology by a “has” relation, the meta node itself is connected to the 4 relations and by this allows analysis on these relations and its assigned concepts.

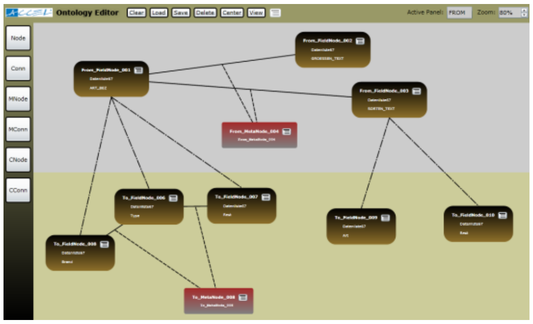


Figure 5. Ontology with meta nodes (for enlarged picture see Appendix)

VI. Creating Results

By the ontology editor, supporting tables, analysis and correction features and meta-nodes, a powerful set of analysis can be performed – which on the other hand can lead to a typical mismatch.

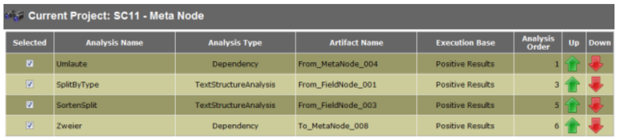
If, for example, one performs a dependency analysis on the first letters of a concept, afterwards splits this concept and then performs a resort, he may receive a different set of results than by first performing the split and resort and then the dependency analysis.

A. *Running and re-running*

To avoid this problem, each analysis should first be run separately to ensure that the results are created as intended. By running and re-running analysis and sequences of analysis, mismatches can be easily identified and a powerful analysis sequence can be developed.

B. *Analysis sequencing*

Once the analysis sequence has been properly set up, one might still wish, to change the analysis sequence (“what happens if…” approach). For this case an analysis sequencer has been implemented to allow not only the resort of values but also a resort of the analysis.



Current Project: SC11 - Meta Node

Selected	Analysis Name	Analysis Type	Artifact Name	Execution Base	Analysis Order	Up	Down
☑	Umlaute	Dependency	From_MetaNode_004	Positive Results	1	⬆	⬇
☑	SplitByType	TextStructureAnalysis	From_FieldNode_001	Positive Results	3	⬆	⬇
☑	SortenSplit	TextStructureAnalysis	From_FieldNode_003	Positive Results	5	⬆	⬇
☑	Zweier	Dependency	To_MetaNode_008	Positive Results	6	⬆	⬇

Figure 6. Sequence of 4 analysis (orig. 6, no. 2 and 4 have been deleted)

VII. Real Life Examples

By now we described approach, procedures and tool on theoretical level. But what is this used for? In this chapter we have collected 4 examples to show the field of application and benefits.

A. *Contract and RFP analysis*

RFPs and contracts are usually longish and full of legal terms.

Lots of companies have made the experience to be closed out from a bidding process just because of formal errors or incomplete proposals. By the analysis features described above it is possible to identify words, elements and rules from a text that describe a must-be, could-be and nice-to-have requirement. This ensures that the company develops a proposal that has all required elements in the right weight and content.

Contracts often have elements that are read over in the beginning (e.g., the usual small print) but lead to severe problems once these elements come to live. If problematic terms from prior contracts are collected in a supporting table, affected elements in new contracts could be identified and a replacement can be proposed based on elements from successful contracts.

B. Standard Service Catalogues

Standard Service Catalogues (SSC) are used to have a valid and efficient basis for contract negotiations. In large companies, SSC usually have a significant size, and sometimes it is hard to identify the correct service.

This opens the door for fraud. If the service contains e.g., to complete some work measured in inch, and the service provider uses centimeter as basis and “forgets” to add the dimension, the customer pays nearly 4 times of the correct price. This can easily brought to light by building a dependency between length and dimension of contract and invoice.

Another typical trick is to invoice extras. As a SSC cannot cover all possible services, extra charges may be applicable for non-SSC services, and these usually are higher than the SSC charges. For services, which seldom occur, some suppliers use “creative” abbreviations. If the SSC e.g., contains “pipe welding DN25 100 mm” and the supplier charges “DN25 weld pp 100” the identity may not be visible at first sight, but is identifiable by a combination of split, resort, dependency and similarity analysis.

C. Article and Material Management

The standard problem in article and material management is always the text. As article and material text are provided by more than one supplier, are manually extended or changed, not only text structure rules are violated but also inconsistencies to other concepts (e.g., between size and weight) may occur. These problems can be resolved by a combination of text structure analysis with supporting tables, splits and dependency analysis.

D. OPTIKON – still running

That last example should be known by all who attended eKNOW 2010. OPTIKON is a research project that tries to resolve the problem of applying a high number of standards on a project performed by a small or medium enterprise (SME). In reality a SME tries to develop a new locomotive. To get the approval for a new locomotive, several hundred different – but in most cases overlapping – standards have to be fulfilled. A SME is overwhelmed, not only by fulfilling but already by reading and understanding all these standards.

The idea of OPTIKON in brief is to identify the rules defined in each standard (using the text input as described in IV.B and the text structure analysis and patterns as described in V.A and V.D), search for duplicates between the rules from different standards (as described in V.E and V.F), limit the rules based on applicable dependencies and not applicable outliers (as described in V.B and V.C) and correct inconsistencies (see V.G and V.H).

After this exercise a minimum set of required rules are available. Now this set can be compared against the internal rules and standards of the SME – still a lot of work but reduced to a minimum must and cleaned from all ballast.

The result will be the – again smaller – set of rules that have to be fulfilled by the SME in addition to its own procedures.

VIII. CONCLUSION

In this paper, we have described how very basic elements from the field of ontologies can be used to generate benefits in industry.

In our experience lots of companies – especially in the SME area – are very conservative in using ontologies as they still assign these methodologies only to the academic world.

Nevertheless if ontologies are used in a fit-for-use approach (“just need a sheet of paper and a pencil”) and then further developed from this hands-on approach to a structured design supported by sufficient analysis features, it can resolve problems with a complexity the user never was able to handle before.

REFERENCES

[1] D. Malzahn: Standard compliant process improvement with ontologies - the OPTIKON project. eKNOW 2010, St. Maarten.

[2] www.levenshtein.de. [retrieved: December 9, 2010]

[3] www.psue.uni-hannover.de/wise2009_2010/apzkett/mat/kap4.pdf. [retrieved: December 9, 2010]

[4] www2.informatik.hu-berlin.de/mac/lehre/SS06/Tag_2_Duplikaterkennung.pdf. [retrieved: December 9, 2010]

APPENDIX

Enlarged versions of Figure 1 and Figure 5

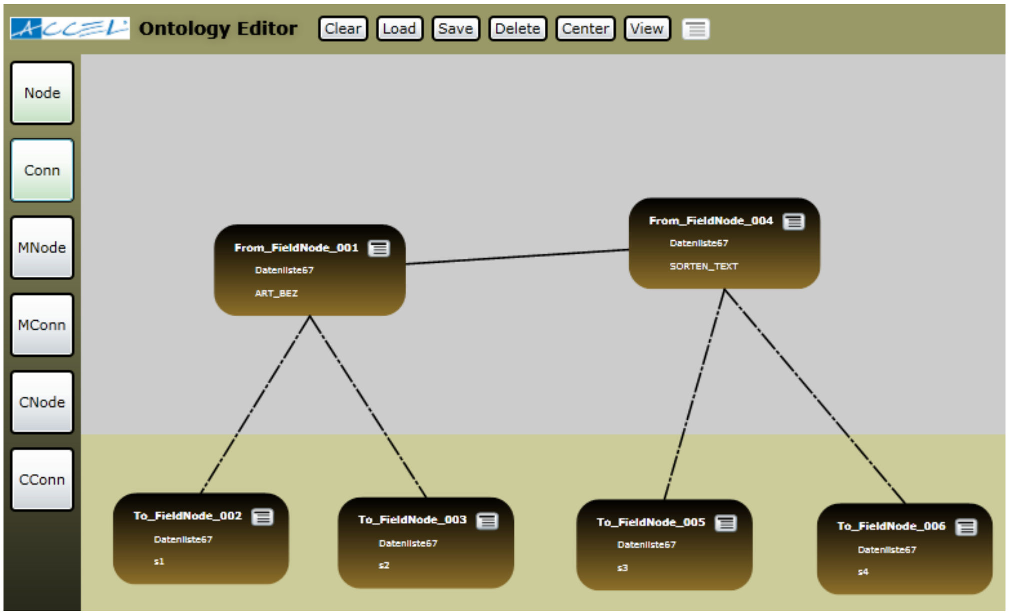
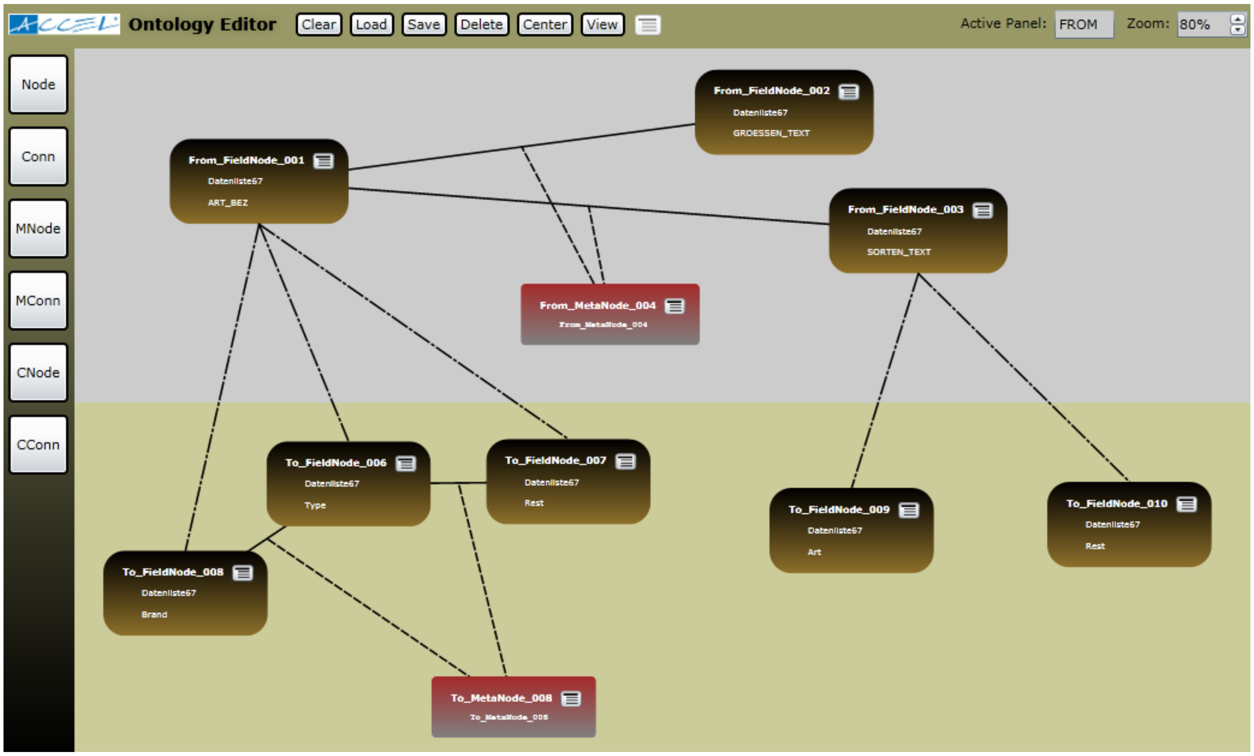


Figure 1. A very basic as-is and to-be ontology



Active Panel: FROM Zoom: 80%

Figure 5 Ontology with meta nodes

BPMN Requirements Specification as Narrative

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait
sabah@alfedaghi.com

Abstract—**The first two phases of the software development process include a requirements analysis stage that demands conceptualization of a "real world domain" and the design stage of the software product. UML-based diagrams are typically used to model systems and make them readable. In this paper we view conceptualization of a piece of reality related to a software system as analogous to a narrative or script created to describe a sequence of events. As an application area, we concentrate on activity diagrams used in BPMN. Examination of typical BPMN representation shows that the resultant picture is fragmented into conceptual gaps and discontinuities. Based on such a perspective, the focus is on maintaining continuity across parts and along the production process of software. To preserve continuity, we propose using the notion of flow as an initial foundation for the conceptualization process.**

Keywords-Activity diagram, BPMN, UML, conceptual model, narrative

I. INTRODUCTION AND MOTIVATION

An information system (IS) should reflect some part of reality and its events. Consequently, building an IS begins by determining requirements as part of a real-world domain. The resulting conceptual picture serves as a guide for the subsequent information system design phase, including a description of the software system under development. According to Peylo [6],

> Requirements engineering is a central part of software projects. It is assumed that two thirds of all errors in software projects are caused by forgotten requirements or mutual misunderstandings in the requirement gathering process. Due to the inherent structure of project planning and the project management process, it is very unlikely that this problem will be solved unless the process itself is changed or we develop tools that possess some intelligence to facilitate the assessment of requirements

Object-oriented methods and languages (e.g., UML) are typically used to describe a software system. Researchers have examined and proposed extending the use of object-oriented software design languages such as UML to apply them at the conceptual level (e.g., [7]). According to Evermann [6], "UML is suitable for conceptual modelling but the modeller must take special care not to confuse software aspects with aspects of the real world being modelled."

In this paper, we concentrate on a specific UML structure, activity diagrams, as applied as a conceptualization tool in BPMN. UML activity diagrams are described as the "flow charts" of object-oriented methodology. The problem with extending object-oriented models and languages is "that such languages [e.g., UML] possess no real-world business or organizational meaning; i.e., it is unclear what the constructs of such languages mean in terms of the business" [6]. The object-oriented IS design domain deals with objects and attributes, while the real-world domain deals with things and properties. According to Storrle and Hausmann [9], in UML, "activity diagrams have always been poorly integrated, lacked expressiveness, and did not have an adequate semantics in UML." With the development of UML 2.0, "several new concepts and notations have been introduced, e.g., exceptions, collection values, streams, loops, and so on" [9].

This paper proposes an alternative approach to specify system requirements. The approach analyzes the relationship between two types of conceptualizations—technical conceptualization and artistic conceptualization—for the purpose of focusing on a main feature of conceptualization: *continuity*.

II. CONCEPTUALIZATION

We view conceptualization as of two types: functional and artistic. *Functional conceptualization* is used for the purpose of representing a piece of reality to be used in building an information system. The resulting artifacts are meant to represent functional requirements. Take for example a UML *use case*, which describes an interaction as a sequence of single steps and events to achieve a specific goal. In this context, there are several representation schemes.

> The meaning (or semantics) of the use case is not represented by the well defined building blocks of the formalism …, but shall constitute itself (helped by various annotations) in the mind of the reader. This approach is quite common but prone to misunderstandings. [6]

ISBN: 978-1-61208-121-2

Furthermore, Peylo [6] states that

> Due to their seeming clarity and formality they are often over-estimated. Nevertheless, they are deceptive with respect to their precision and expressiveness. Their main limitations are:
> 1. Weak and not well defined semantics of relations.
> 2. The expressiveness of graphical representation schemes is limited per se to a fragment of first order logic
> 3. Generally, it is not possible to decide by the study of a use case whether the process flow may lead to the desired result (i.e. the system output may be achieved, given the set of input).

Artistic conceptualization is also generated for the purpose of representing a part of reality. It can be exemplified by narratives, scripts of movies, and comic books.

Both types of conceptualization are strongly founded on language. Their orientations are different, as shown in Fig. 1. Artistic conceptualization captures reality but seeks to release its content in an expanded universe of meanings and interpretations. Functional conceptualization seeks precision in releasing its content by narrowing its meaning and interpretation. An important aspect of both types of conceptualization is *continuity,* as described in the next section.

A notion related to artistic conceptualization is that of "operation concept," which includes concept analysis. Concept analysis is an overall "system development process" for analyzing an operational environment and characterizes a proposed system from the user's perspective.

> [An operation concept] document should, in contrast to a requirements specification, be written in narrative prose, using the language and terminology of the users' application domain. It should be organized so as to tell a story, and should make use of visual forms (diagrams, illustrations, graphs, etc.) whenever possible [5].

However, this does not focus on the notion of *flow* (a fundamental concept in our approach that will described later) even through it recommends "scenarios [that] are specified by recording, in a step-by-step manner, the sequences of actions and interactions between a user and the system" [5].

III. CONTINUITY

In a system, continuity indicates uninterrupted connection and succession. In the production of film and television, a script supervisor is concerned with maintaining *continuity* across shots and along the production process. In comic books, *continuity* means contiguous events "in the same universe."

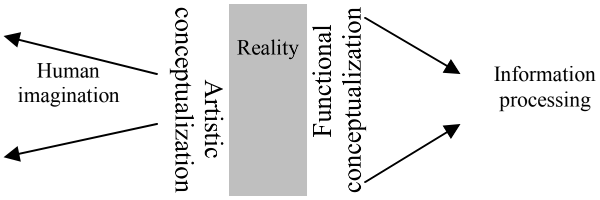


Figure 1. Orientations of artistic and functional conceptualizations.

In business, the notion of continuity/security arises when planning for permanence of critical business processes in case of security failure. It is a notion related to survivability, load balancing, and redundancy.

In beginning mathematics a function is continuous if we can draw its graph without taking the pencil off the page. A *discontinuity* is a point where a function is not continuous.

Ivic [4] defines *discontinuity* as "the lack of … logical sequence." According to Webster's New World College Dictionary [10], discontinuity means "a lack of continuity or *logical sequence*, or a gap or a break… this could mean a break in the chronological sequence, or a very fragmented structure in poetry." Discontinuity is an undesirable feature in literature. "Discontinuity in a novel interrupts the *flow* of the story, ..." [10, italics added].

In architectural design, continuity is "the measurement of the *completeness* of the sidewalk system with avoidance of *gaps*… the pedestrian sidewalk appears as a single entity within a major activity area or public open space" [11, italics added].

In a software system, *discontinuity* may be a positive feature for security. "System discontinuity emphasizes security over compatibility by removing those constructs in our system software which lead to security holes in applications" [12]. Such strategy removes parts of the interfaces "both of programming languages and operating systems which have proven to engender the greatest number of security holes." Such a proposal assumes completeness. In analogy, to secure a physical territory, subterritories can be disconnected; however, the interior of each piece of territory should be completely known (e.g., surveyed).

A conceptualization of reality needs a type of continuity: *logically sequential progression*. This can be thought of as reflecting the Aristotelian notion of organic unity, where each component of a task is a necessary part of a whole.

Continuity is a necessary feature for designers. After producing a conceptual representation, designers will seek connections through temporal continuity, causality, or some commonality such as presence in the same sphere.

In general, the notion of continuity is a phenomenon that involves a gradual transition without abrupt changes or discontinuities. We view it as the property of connectedness of conceptual space of events. When a conceptualization seems fragmentary, we look at the represented world. Is there an underlying represented "reality" that can be pieced together? Are there missing entities or connections? Are there discontinuities between spheres? Do conceptual parts have gaps that can't be assertively filled?

We show the importance of continuity through scrutinizing BPMN activity diagramming. To provide opportunities to contrast continuous and discontinuous representations, we next review a flow-based conceptualization that can be used for modeling activities.

IV. FLOWTHING MODEL (FM)

A flow model is a uniform method for representing things that "flow," i.e., things that are exchanged, processed, created, transferred, and communicated [1, 2]. "Things that flow", called flowthings, include information, materials (e.g., in manufacturing), and money. To simplify this review of FM, we introduce the model in terms of information flow. Information occurs in five states: transferred, received, processed, created, and released, as illustrated in Fig. 2. Here, we view a "state of information" in the sense of properties; for example, water occurs in nature in the states of liquid, solid, and gas.

Fig. 2 also represents a transition graph, called a flowsystem, with five information states and arrows representing flows among these states. Information can also be stored, copied, destroyed, used, etc., but these are secondary states of information in any of the five generic states. In Fig. 2, flows are denoted by solid arrows. Flows may *trigger* other types of flow, denoted by dashed arrows, as will be discussed.

The environment in which information exists is called its sphere (e.g., computer, human mind, organization information system, department information system). The flowsystem is reusable because a copy of it is assigned to each entity (e.g., software system, vendor, and user). An entity may have multiple flowsystems, each with its own flowsystem. It is possible to have flowsystems of different flowthings: requests, invoices, plans, and actions. These are, like information, flowthings that can be received, processed, created, released, and transferred.

A flowsystem may not necessarily include all states, for example, conceptualization of a physical airport can model the flow of passengers: arriving (received), processed (e.g., passports examined), released (waiting to board), and transferred (to planes); however, airports do not create passengers (ignoring the possibility of an emergency where a baby is born in the airport). In this case, the flowsystem of the airport includes only passenger states of received (arrival), processed (e.g., passports), released (waiting for boarding), and transferred (on the plane).

As we mentioned previously, we view a system as the environment in which information exists, called its sphere. A system is also viewed as a complex of flowsystems.

The states shown in Fig. 2 are exclusive in the sense that if information is in one state, it is not in any of the other four states. Consider a piece of information x in the possession of a hospital. Then, x is in the possession of the hospital and can be in only one of the following states:

1. x has just been collected (*received*) from some source, e.g., patient, friend, or agency, and stored in the hospital record waiting to be used. It is received (row) information that has not been processed by the hospital.
2. x has been *processed* in some way, converted to another form (e.g., digital), translated, compressed, etc. In addition, it may be stored in the hospital information system as processed data waiting for some use.
3. x has actually been *created* in the hospital as the result of doctors' diagnoses, lab tests, produced by processing current information (e.g., data mining), and so forth. Thus, x is in the possession of the hospital as created data to be used.

If a piece of information is copied, then the new piece of information is a different instance of a flowthing (e.g., one is stored, and one is transferred).

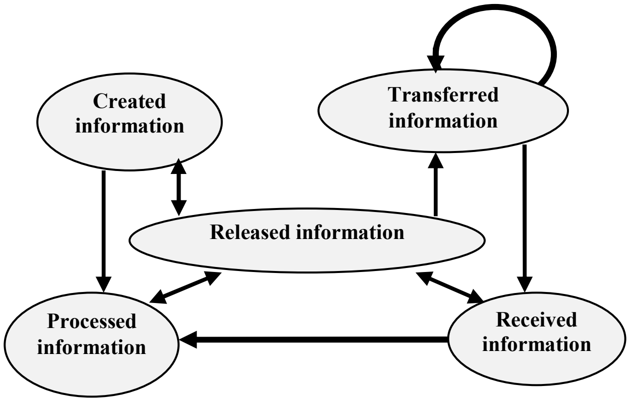


Figure 2. State transition diagram of FM with possible triggering mechanism.

4. x is being released from the hospital information sphere. It is designated as released information ready for transfer (e.g., sent via DHL). In an analogy of a factory environment, x would represent materials designated as ready to ship outside the factory. They may actually be stored for some period waiting to be transported; nevertheless, their designation as "for export" keeps them in such a state.

5. x is in a transferred state, i.e., it is being transferred between two information spheres. It has left the released state and will enter the received state, where it will become received information in the new information sphere.

It is not possible for processed information to directly become received information in the same flowsystem. Processed information can become received information in another flowsystem by first becoming released information, then transferred information, in order to arrive at (be received by) another flowsystem.

Consider the seller and buyer information spheres shown in Fig. 3. Each contains two flowsystems: one for the flow of orders, and the other for the flow of invoices. In the seller's infosphere, processing of an order triggers (circle 3) the creation of an invoice in the seller's information sphere, thus initiating the flow of invoices.

The reflexive arrow of the transfer state shown in Fig. 2 (above) denotes flow from the transfer state of one flowsystem to the transfer state of another.

In Fig. 3, the Buyer creates an Order that flows by being released and is then transferred to the Seller. The "transfer components" of the Buyer and the Seller can be viewed as their transmission subsystems, while the arrow between them represents the actual transmission channel.

V. BPMN ACTIVITY DIAGRAM

Business Process Modeling Notation (BPMN) is popular in some communities of practice and "in some cases may be locally mandated" [8]. Therefore, it is useful to utilize it as an area where different activity conceptualizations are compared.

When developing a business system, it is essential to first produce a general conceptual description of *activities*. The activities pose scenarios that represent the circumstances of events. The resultant description is a model of overall activities, subactivities, and connections among them. This conceptualization liberates designers to produce neutral specifications not oriented to any actual current methodology of conducting business. It also represents a common understanding of system operations shared by technical and nontechnical individuals involved in the project.

BPMN shows activities within swimlanes, which represent different performers as nodes in the Business Node Connection Model. Fig. 4 illustrates the basic form of a BPMN diagram, in the context of a travel planning activity [8 - Citizant Corp.]. Fig. 5 shows the corresponding FM representation. According to Sowell [8],

> This all-in-one notation can be very helpful and time-saving when the architecture in question is an As-Is architecture, because *all the relevant information* is known, and merely needs to be captured. [Italics added]

We claim that the activity diagram shown in Fig. 4 exhibits a fragmented conceptualization of reality. The workflow description items form a narrative that is created to describe a sequence of events. Consequently, we go one item at a time, as follows. We assume that the software designer is the reader of such a narrative.

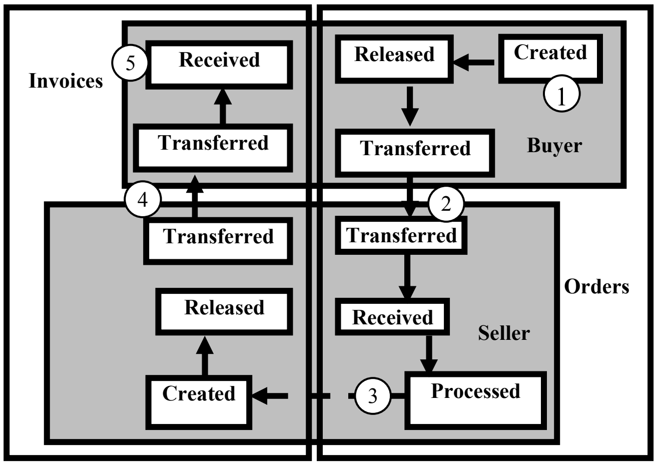


Figure 3. Order flow triggers invoice flow.

Consider the following scenario in Fig. 4.

Travel agent: Research Travel Options
Traveler: Select Itinerary
Travel agent: Make Reservation
Traveler: Submit Payment
Travel agent: Confirm reservation
Traveler: Verify Itinerary

The arrows in the figure seem to indicate control flow. The semantics involved are as follows:

The travel agent researches travel options,
the traveler selects an itinerary,
the travel agent makes the reservations,
the traveler submits payment,
the travel agent confirms the reservation,
the traveler verifies the itinerary.

Here we see a discontinuity. For example, in the sequence:
[the travel agent confirms the reservation →
the traveler verifies the itinerary] the events seem to jump.

A corresponding scenario with continuity would be as follows:

The travel agent researches travel options,
the search by the travel agent produces a list of options,
the travel agent sends the list to the traveler,
the traveler selects an itinerary from the list,

ISBN: 978-1-61208-121-2

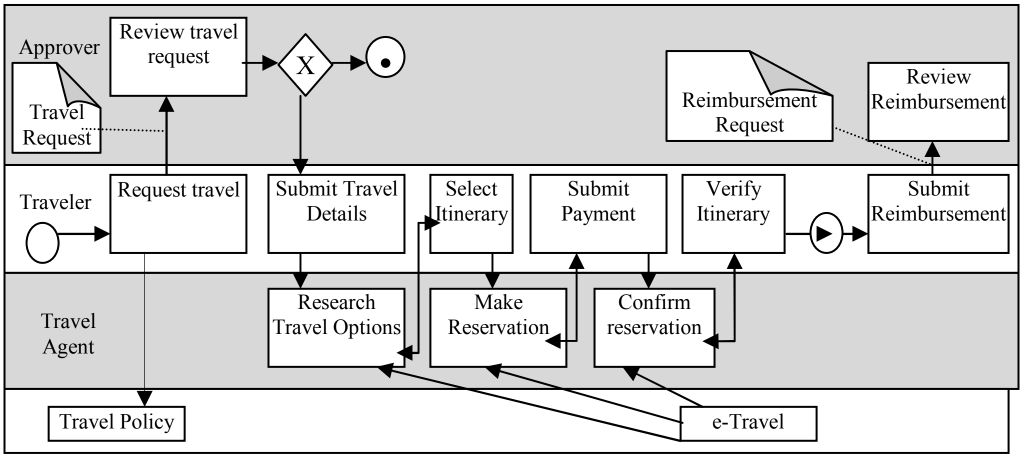


Figure 4. A Simple BPMN Diagram [8].

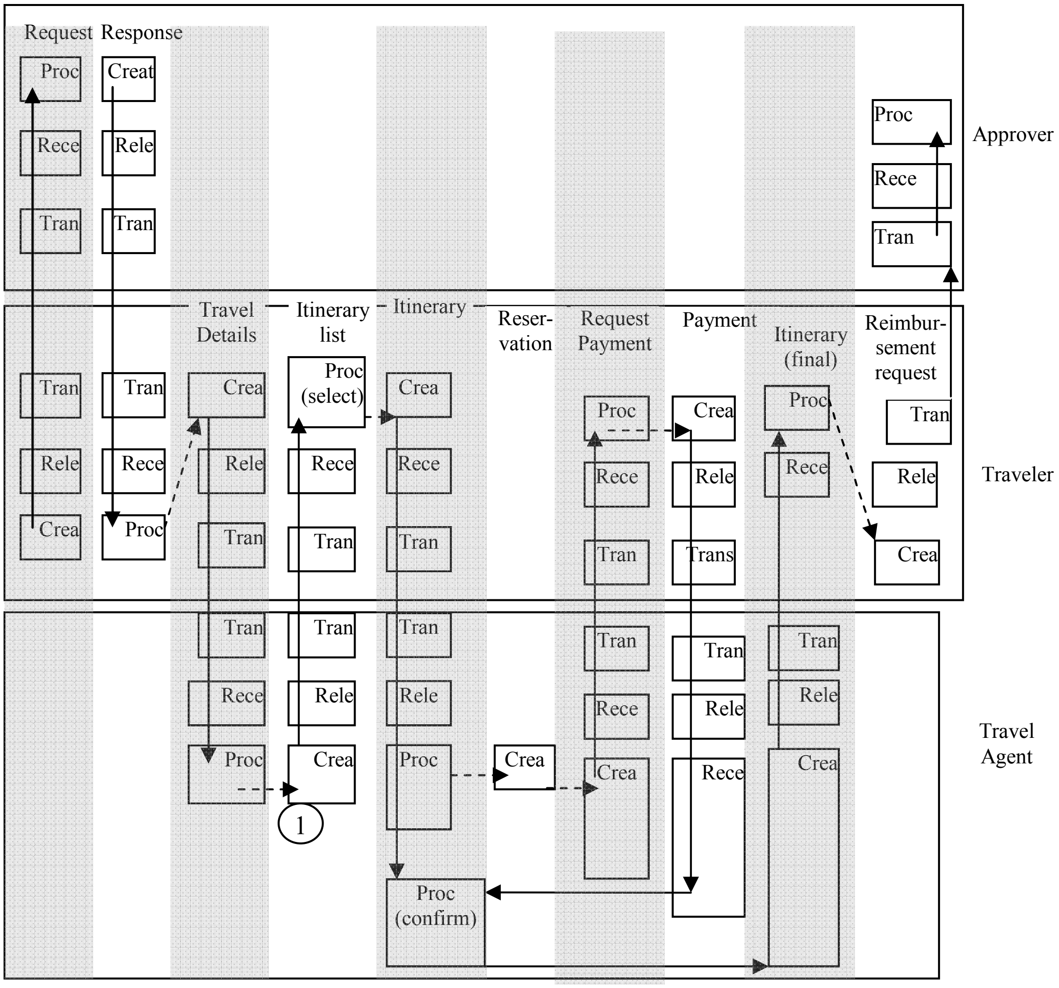


Figure 5. Flow-based representation of a Simple BPMN Diagram

the traveler sends the itinerary to the travel agent,
the travel agent makes the reservation,
the travel agent issues a payment invoice,
the travel agent sends the invoice to the traveler,
the traveler receives the invoice,
the traveler makes payment(e.g., money order)
the traveler sends payment to the travel agent
the travel agent receives payment from the traveler,
the travel agent confirms the reservation,
the travel agent sends the itinerary to the traveler,
the travel agent confirms the reservation, then
the traveler verifies the reservation.

Fig. 5 reflects such continuity. Starting at circle 1, the travel agent creates an itinerary that flows to the traveler, which triggers him/her to select a single option (itinerary) that flows back to the travel agent, who processes it and (1) makes a reservation, and (2) creates an invoice. The invoice is sent to the traveler, who makes (creates) payment, which arrives at the travel agent. The travel agent confirms the reservation, and sends the final itinerary to the traveler. Upon receiving the itinerary, the traveler processes it to verify it.

This flow-based description is similar to a comic book, where a stream of events flows in a continuous fashion. Flowthings such as requests, lists, and invoices flow like a ping pong ball between players.

VI. WITH WORKFLOW DESCRIPTION

"Use case" as a modeling tool provides a software-independent description of the processes to be automated.

> The IT team must have descriptions of the business that allow team members to make informed decisions, including an unambiguous specification of the business process that details relevant value and cost factors. Business use cases are documented via specifications that consist of both textual workflow descriptions and one or more Unified Modeling Language (UML) activity diagrams. [3]

Consider Fig. 6, which provides an example of a business use case specification [3]. The activity diagram provides a pictorial representation of the workflow structure described in the following business use case text. [3] gives the corresponding workflow description for Fig. 6. For lack of space, we discuss the first three steps as follows:

• The Customer Sales Interface initializes contact.

• If the Customer Sales Interface determines that initial opportunity work is complete, then the Customer Sales Interface sends a proposal request to the Proposal Owner.

• Otherwise the Customer Sales Interface searches for alternatives. [3]

As stated previously, the workflow description items form a narrative that describes a sequence of events. We assume that the software designer is the reader of such a narrative.

• The Customer Sales Interface (CSI) initializes contact.

From such a description, implicitly (from the name), we understand that there is a customer. Contact denotes communication, thus, it seems that the designer would understand that CSI creates something (e.g., a message) and then executes the contact. To maintain continuity and completeness in the initial step, we must explicitly state that something is created, as follows.

CSI creates an offer and communicates it to customer.

Here we ignore the issue of what type of information is involved in such a creation.

• If the Customer Sales Interface determines that initial opportunity work is complete, then the Customer Sales Interface sends a proposal request to the Proposal Owner (PO).

This scenario includes missing pieces. How does the designer understand that the "determination" is the result of receiving some type of communication from the customer? It is possible that the designer thinks that embedding some type of information about communication with a customer is unnecessary since the determination is based on informal contact. It is highly improbable that contact with a customer is non-recorded informal contact. We can rewrite this as follows.

CSI receives a response from the customer, processes the response, then If CSI determines that initial opportunity work is complete, CSI sends a proposal request to the PO.

The whole process can be described as flows of offers, responses, and requests as partially shown in Fig. 7.

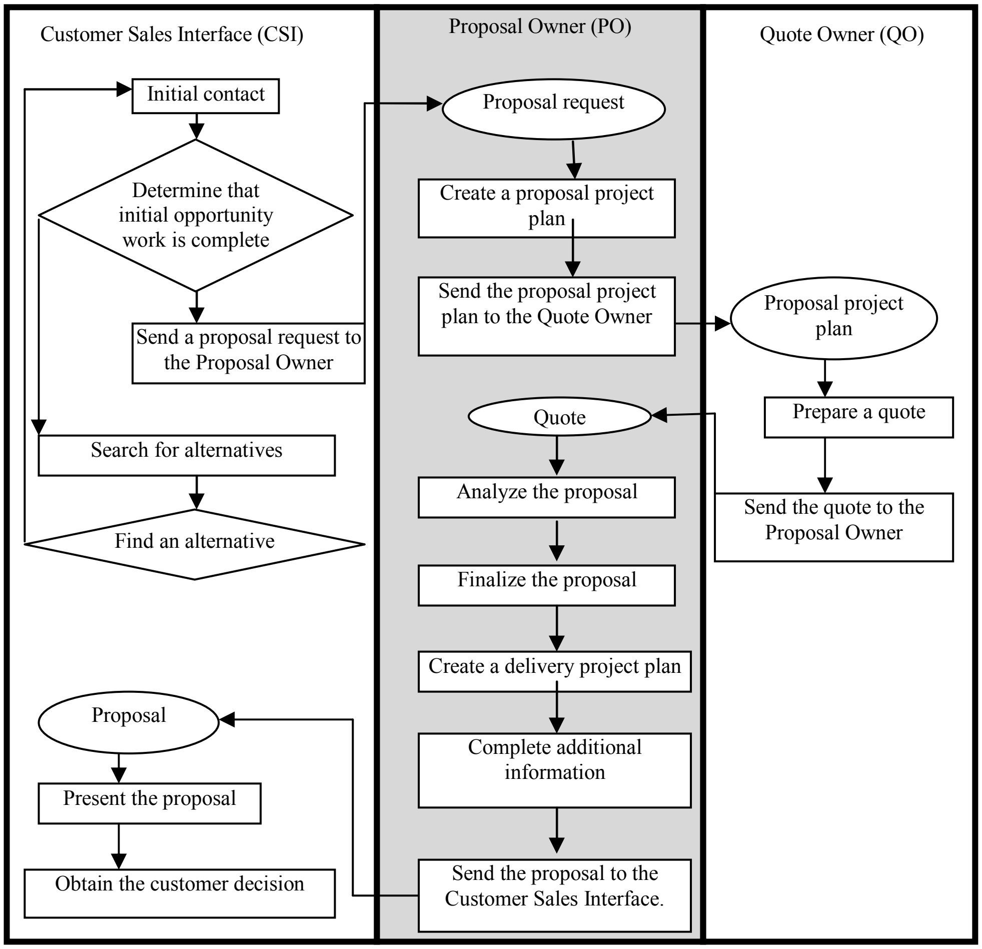

Customer Sales Interface (CSI)
Proposal Owner (PO)
Quote Owner (QO)
Initial contact
Determine that initial opportunity work is complete
Send a proposal request to the Proposal Owner
Search for alternatives
Find an alternative
Proposal
Present the proposal
Obtain the customer decision
Proposal request
Create a proposal project plan
Send the proposal project plan to the Quote Owner
Quote
Analyze the proposal
Finalize the proposal
Create a delivery project plan
Complete additional information
Send the proposal to the Customer Sales Interface.
Proposal project plan
Prepare a quote
Send the quote to the Proposal Owner

Figure 6. Example of a business use case specification (From [3]).

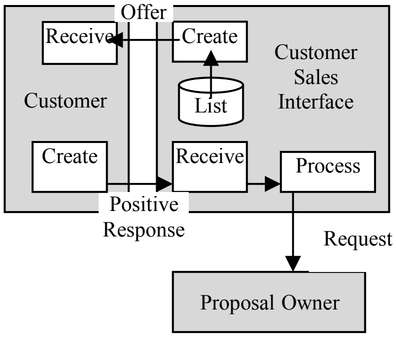


Figure 7. Non-programming conceptualization of part of the example.

Notice that a general conceptualization (shown partially for lack of space) in Fig. 7 reflects a "forest-level" of flows in the piece of reality being abstracted. There is the flow originating from CSI to customer, and another flow originating from customer that may reach PO. The *need* for flow has been expressed previously in a discussion of Peylo's [6] "flow of action" in scripts, and "flow of the story" in [11].

Notice also the general level of conceptual mapping in FM. When designing a city, the designer does not specify at intersections that green means go and red means stop. These details (types of processes in FM) come at a lower level of abstraction. Thus, it is not necessary, in our example, to specify at this level, that "If the Customer Sales Interface determines that initial opportunity work is complete, then the Customer Sales Interface sends a proposal request to the Proposal Owner." It is sufficient to indicate at this point that:

- Responses are received from customers, processed, and according to this process a request is sent to Proposal Owner.

FM description draws a conceptual topology of flows of data, leaving the interior of the process (e.g., specification of decision criteria) to a later stage. Accordingly, the designer can visualize the total procedure as:

For list of customers
Send an offer
Receive a response
Process the response
According to the results of processing, send/do not send a request for Proposal Owner

Note that there is no "if" statement in this procedure, because "if" triggers specification of the criteria for a decision.

Additionally, going back to the narrative of workflow of [3], we see the following.

• *The Quote Owner (QO) prepares a quote.*
• *The Quote Owner sends the quote to the Proposal Owner.*

Here one wonders why "prepare" is used, instead of "create," as used previously by [3]. "Create" is more suitable because it is a flow-oriented term: QO creates (originates) quotes that flow to PO.

In Fig. 6, there are odd arrows (dataflow? control flow?) from a process to an object, such as the arrow from "Send the proposal project plan" to the "Quote Owner", and the arrow from "Send the quote to the Proposal Owner" to "a quote".

We stop here reviewing the rest of the workflow and activity diagram because it is clear at this point that such a description is "narrative-wise", is a fragmented conceptualization that is filled with gaps, and discontinuities.

Finally, we note the uncontrollable use of many verbs: "initializes", "determines", "searches", "finds", "sends", "prepares", "creates", "analyzes, "finalizes", "completes", "presents", and "obtains". This style of specifying flow among processes is a frail feature in any good "conceptual narrative". In contrast, FM uses only five flow-oriented operations: receive, process, create, release, and transfer.

Clearly, we are not introducing a completely new methodology for specifying requirements; rather we describe a general approach that emphasizes flow and continuity of requirements description.

VII. CONCLUSION

This paper introduces the concept that a piece of reality related to a software system can be conceptualized analogous to a narrative or script created to describe a sequence of events. This is demonstrated by applying it to activity diagrams used in BPMN utilizing flow-based model. The resultant description maintains continuity across parts and along the production process of software. Further research would explore applying the concept to other software diagramming tools.

REFERENCES

[1] S. Al-Fedaghi, "Conceptualization of business processes," IEEE Asia-Pacific Services Computing Conference (IEEE APSCC 2009), Dec 7-11, 2009, Biopolis, Singapore.

[2] S. Fedaghi, "Scrutinizing UML activity diagrams," 17th International Conference on Information Systems Development (ISD2008), Paphos, Cyprus, August 25-27, pp. 59-67, 2008.

[3] A. Frankl, "Validated requirements from business use cases and the Rational Unified Process," IBM, 15 Aug 2007, accessed June, 2010. http://www.modernanalyst.com/Resources/Articles/tabid/115/articleType/ArticleView/articleId/52/Validated-requirements-from-business-use-cases-and-the-Rational-Unified-Process.aspx

[4] C. Ivic, "Review of discontinuities: new essays on Renaissance literature and criticism," Early Modern Literary Studies 5, 2, September, 1999, 8.1-6, accessed June 2010. http://purl.oclc.org/emls/05-2/ivicrev.htm

[5] R. E. Fairley, R. H. Thayer, and P. Bjorke, "The concept of operations: the bridge from operational requirements to technical specifications," Proceedings IEEE International Conference on Requirements Engineering , 18-21 April 1994, Colorado Springs.

[6] C. Peylo, "On restaurants and requirements: how requirements engineering may be facilitated by scripts," The 4th Workshop on Knowledge Engineering and Software Engineering (KESE 2008), accessed June, 2010. http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-425/paper7.pdf

[7] J. Evermann, and Y. Wand, "Towards ontologically based semantics for UML constructs," In: Kunii, H, Jajodia, S, and Solvberg, A (eds.), Proceedings of the 20th International Conference on Conceptual Modeling, pp. 354-367, 2001, Yokohama, Japan.

[8] K. Sowell, "Creating and Presenting Activity Models," SowellEAC Blog, Accessed, December, 2009. http://sowelleac.com/Creating_and_Presenting_Activity_Models.pdf

[9] H. Storrle, and J. H. Hausmann, "Towards a formal semantics of UML 2.0 activities," German Software Engineering Conference, pp. 117-128, 2005. http://wwwcs.uni-paderborn.de/cs/ag-engels/Papers/2005/SE2005-Stoerrle-Hausmann-ActivityDiagrams.pdf.

[10] ENH241, "American Literature before 1860, Discontinuity," Accessed June, 2010. http://enh241.wetpaint.com/

[11] Kansas City, "Walkability Plan. 32. Measuring Walkability," Accessed June, 2010, . http://www.kcmo.org/idc/idcplg?IdcService=GET_FILE&dID=26423&dDocName=019904

[12] J. A. Solworth, "Robustly secure computer systems: a new security paradigm of system discontinuity," Proceedings of the 2007 Workshop on New Security Paradigms, 2007.

Run-time Adaptable Business Process Decentralization

Faramarz Safi Esfahani

Department of Software Engineering,
Islamic Azad University, Najaf Abad Branch,
Esfahan, Iran.
fsafi@iaun.ac.ir

Masrah Azrifah Azmi Murad,
Md. Nasir Sulaiman, Nur Izura Udzir
Faculty of Computer Science and Information Technology
University of Putra Malaysia, 43400, Serdang,
Selangor, Malaysia.
{masrah, nasir, izura}@fsktm.upm.edu.my

Abstract **- BPEL specified business processes in the Service Oriented Architecture (SOA) are executed by non-scalable centralized orchestration engines. In order to resolve scalability issues, the centralized engines are clustered, which is not a final solution either. Alternatively, several decentralized orchestration engines are being emerged with the purpose of decentralizing a BPEL process into fragments, statically. Fully decentralization of a process into its building activities is an example of static fragmentation methods. The fragments are then encapsulated into run-time components such as agents. There are a number of attitudes towards workflow decentralization; however, only a few of them consider the adaptability of produced fragments with a run-time environment. The run-time adaptability can be studied from different aspects such as the proportionality of workflow fragments with number of machines dedicated to a workflow engine or runtime circumstances such as available bandwidth. In our opinion, the SOA suffers from the lack of decentralization adaptability with run-time environments in the orchestration layer. It demands the mapping of run-time circumstances to a suitable fragmentation model. In this paper, a mapping algorithm is presented, which is based on the number of machines and available bandwidth. Evaluation of the presented algorithm for adaptable decentralization demonstrates an improvement of the bandwidth usage compared to a fully decentralized process.**

Keywords-Adaptive Systems; Service Oriented Architecture; Distributed Orchestrate Engine; Self-* Systems; BPEL; Mobile Agents;

I. Introduction

Business processes might be very large, geographic location dependent, long running, carrying a vast number of calculations, manipulating a huge amount of data and will eventually be realized as thousands of concurrent process instances. Such workflows might be found in different areas of industry and even technology. Authors in [1], refer to the applications of business processes in industries such as chain management, online retail, or health care to consist of complex interactions among a large set of geographically distributed services deployed and maintained by various organizations. In addition, an electronic manufacturer is also reported that employs business processes to conduct its operations including component stocking, manufacturing, warehouse, order management and sales forecasting. There exist geographically distributed parties such as a number of suppliers, several organizational departments, a dozen of sales centers, and many retailers. Requests for such processes from different parties all together naturally result in creating thousands of concurrent executing instances. Such number of concurrent requests is a natural fit to this paper. In addition, new software paradigms introduced in the Cloud computing such as software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) are targeted to receive a huge number of requests. Particularly, in the case of orchestrate engine as a service, a huge number of workflow instances might be deployed and requested from various clients all around the world. In order to handle the requests, different number of machines and also resources must be employed. This paper proposes an adaptable and distributed workflow engine, which tackles such an ever-changing environment.

According to the SOA stack [2], business logic layer consists of orchestration and choreography layers. The choreography layer is intrinsically distributed to several distinct workflows communicating with each other and normally run on different workflow engines, whereas the orchestration layer is workflow engine centric. Indeed, a single engine [3] is usually applied to execute a business process and scalability is naturally addressed by replicating orchestration engines, which is not a final solution for scalability problems of centralized engines, entirely [1, 4, 5]. The decentralization of business processes has been introduced as an alternative solution, which is currently based on system analyst and designers' opinion and is carried on at design time, without paying attention to the fact that ever-changing run-time environment raises special requirements on which there is no information at design time.

From this paper point of view and according to [2], business process decentralization methods can be studied from three aspects including fragmentation, enactment and adaptability. A number of decentralized workflow engines have emerged to support these aspects of decentralization. The most challenging area is adaptability, which means the ability of system to reconfigure its component to refrain from or lessen system bottlenecks such as throughput, response time and bandwidth usage. In order to achieve adaptability, the system must be able to react to run-time circumstances and reconfigure itself. A fully process decentralization (FPD) method is applied by [1, 4, 6-10], which decentralizes a business process to activity level fragments. The fragments are encapsulated in run-time components and a third-party middleware is applied to support the communication among fragments. Adaptability also comes about by locating/relocating the run-time components based on system conditions. The FPD negatively produces a number of fragments, which their message passing and resource usage will eventually result in swamping the run-time environment. The main reason is that the fragments are statically produced without considering run-time circumstances. In contrast, there are several dynamic fragmentation methods [11-13], which produce fragments at run-time without considering the run-time circumstances. Thus, the produced fragments may cause violating system thresholds.

In our opinion, the mentioned methods suffer from the following problems. 1) Lack of dynamic criteria to workflow decentralization. 2) Improper selection of activities in decentralized process fragments. On one hand, encapsulating each activity in one run-time component provides a high number of components, along with a plenty of message passing and will eventually increase bandwidth usage. It also results in high response time due to a huge amount of message passing and most importantly low system throughput due to high resource consumption. On the other hand, encapsulating coarser fragments based on static criteria will result in less system flexibility as well as adaptability with run-time environment.

Having an abstract layer in the SOA architecture to realize the adaptability of decentralization with run-time environment can be a solution for the mentioned problems. This layer may seem to be an overhead for executing processes; however, it is a tradeoff among different aspects of system. The negative effects of this layer can also be mitigated by creating the fragments in advance, managing workflow states and etc. There have been experiments [14] that show the gained advantages is eye-catching enough which is a good motivation for presenting this abstract layer. Indeed, the main objective of this work is presenting and implementing the idea of adaptable business process decentralization based on current run-time circumstances. In fact, the current system condition is mapped to a suitable decentralization method. In addition, the contributions of this work are: 1) introducing the idea of run-time adaptable business process decentralization. 2) Presenting a bandwidth adaptable workflow decentralization method.

It is also worth mentioning that this paper focuses on block-structured business processes. In addition, several aspects of workflow management systems are not included in this work such as governance of workflow fragments, managing run-time state of workflow fragments, run-time workflow reconfiguration, transaction, exception handling and sharing workflow fragments interior variables. These are normally implemented by a distributed middleware [11-13] or from dynamic process decentralization view; they demand more attention in future work as well.

II. Background and Related Work

Open World Software Paradigm: Baresi et al., in [15], open a new view towards software development by introducing the idea of open-world software paradigm, which has attracted much attention nowadays. According to this idea, in the open-world paradigm, software is executed in an ever-changing environment; therefore, static design time metrics will not be responsive at run-time. Although the run-time environment changes continuously, it is the software itself, which has to be self-healed and self-adapted to keep the whole system in a safe side. Generally, changes in run-time environment are not predictable at design time; therefore, evolving software at run-time is necessary. Software thus needs to continuously and automatically adapt itself and react to the changes. Systems will need to operate correctly despite of unexpected changes in factors such as environmental conditions, user requirements, technology, legal regulations, and market opportunities. This work brings an example of applying open-world idea to service oriented applications.

A few research works has been performed on self-adapting and self-healing of the service-oriented applications, especially from service composition point of view. However, none of them considers self-adapting of business process decentralization with run-time environment. This paper draws the idea of open-world paradigm into business process composition from decentralization point of view. A decentralized workflow engine is required to be adequately flexible, dynamic and adaptive to handle the changes by providing adaptable fragments. The focus of this work is on implementing a decentralized workflow engine, which creates adaptable fragments from a business process based on run-time environment feedbacks. Adaptability comes about in terms of first) number of dedicated machines to a workflow engine; second) the available bandwidth of media, which connect the workflow engine machines. Based on these adaptability aspects, a bandwidth adaptable algorithm is also presented to choose a suitable decentralization method as well.

Adaptability of Decentralization Based on Workflow Circumstances: Dartflow project [11] has shown an usage of mobile agents in distributed workflow execution. In Dartflow, the workflow model is fragmented dynamically, and the partitions are carried by mobile agents and sent to different sites, which are responsible for them. This work establishes good points for dynamicity of error handling and data sharing among agents in a dynamic workflow system; however, it focuses on the system architecture and does not detail the fragmentation model, adequately.

An abstract and conceptual dynamic workflow fragmentation method is shown by [12], which applies the Petri net formalism. The presented method partitions the centralized process into several fragments step by step, while the process is executed. The created fragments are able to migrate to proper servers, where tasks are performed and new fragments are created and forwarded to other servers (i.e., using mobile agents) to be executed. Dynamism in decentralization is a prominent aspect of this work. The fragmentation method of this work is different from our work in that it considers workflow run-time condition to decentralize a business process, while run-time environment circumstances are applied for the same purpose in this paper. Nonetheless, this method is similar to our work in that the fragments must be prepared beforehand in a fragment pool or must be built on the fly at runtime.

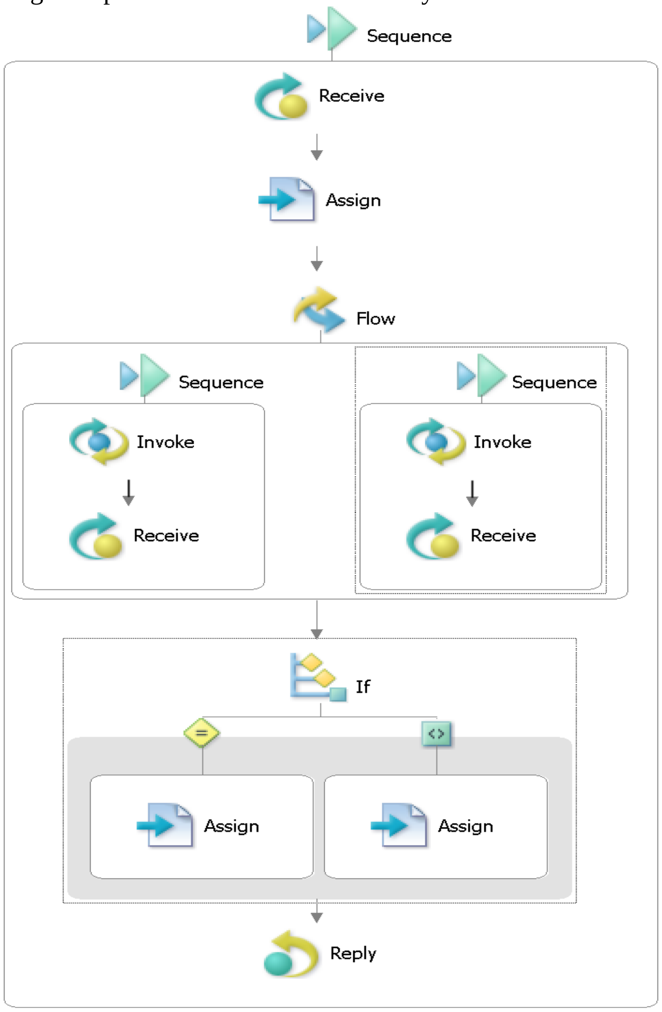


Figure 1: BPEL View of Loan Application Process

In [13], a decentralized workflow model is presented for inter-organizational workflow decentralization, where inter-task dependencies are enforced without requiring to have a centralized WFMS. This work is different from our work in that it considers a criterion to decentralize a business process from inter-organizational point of view which is fitted to choreography layer of service oriented architecture. The produced fragments can be considered as inputs of our work for dynamic decentralization. This work is also an inter-organizational version of the research paper [12], which partitions the workflows on the fly. It is also different from our work in that a workflow is partitioned based on workflow run-time conditions, while this paper is targeted to use run-time environment circumstances for decentralization purposes.

In [16], the ever-changing legislation of governments, customers' needs and other changes in environment of business processes are introduced as the main reasons of implementing various forms of business process applications in different organizations. This research work looks for a way of providing highest level of flexibility as well as adaptability to the changes in run-time environment. The decentralization methods introduced in this current paper require flexible architectures to support dynamic fragmentation, which execute different execution forms of a business process at runtime.

Adaptability of Decentralization Based on Run-time Circumstances: This study is also an extension of our previously published papers, which were totally on introducing dynamic criteria for business process decentralization. In [17, 18], mere idea and motivations of using a mining method for intelligent process decentralization (IPD) were introduced and the improvement of only response time was *mathematically* shown for several sample BPEL processes. Moreover, [19] showed an SLA driven aspect of the IPD as well. Furthermore, hierarchical process decentralization criterion (HPD) and its composition with the IPD for two different case studies were shown in [14, 20]. In this current research study, the HPD method is considered as a decentralization method, which decentralizes a business process based on the hierarchy of activities. For instance in $level_0$ of the process tree, the whole process is considered as a fragment that is equal to a Centralized process. In the $level_{n-1}$, which is the last level of the process tree, each process activity is considered as an individual fragment and it is analogous to the FPD method. The middle layers of the process tree provide coarser fragments. Based on the level of decentralization, different numbers of fragments are produced. This paper presents a method for process decentralization, which applies the number of workflow engine machines and available bandwidth to choose a suitable level of decentralization in the process tree.

III. HPD Decentralization of Loan Process

Unfortunately, there is no standard business process for benchmarking BPEL processes. Nonetheless, a loan application business process has been applied in [1, 10, 21, 22] that also fits our research. The loan process illustrated in Figure 1 is decentralized based on the HPD decentralization approach as shown in Figure 2. It makes us able to study several simple and structured BPEL activities together. The loan process consults with two external web services, which send a credit report for the loan applicant. In order to refrain from approbating risky loans, the loan request is accepted, when both web services confirm the applicant's credit. According to the HPD decentralization method, the loan process can be decentralized based on the levels of the process tree. The $level_0$ or HPD0 contains only one fragment, which is equal to the Centralized model. The $level_1$ which is analogous to HPD1 provides six fragments, HPD2 in the $level_3$ decentralizes the process to ten fragments and finally HPD3, which is the finest fragmentation model and is also called the FPD, fully decentralizes the loan process into sixteen fragments each of which contains only one activity.

ISBN: 978-1-61208-121-2

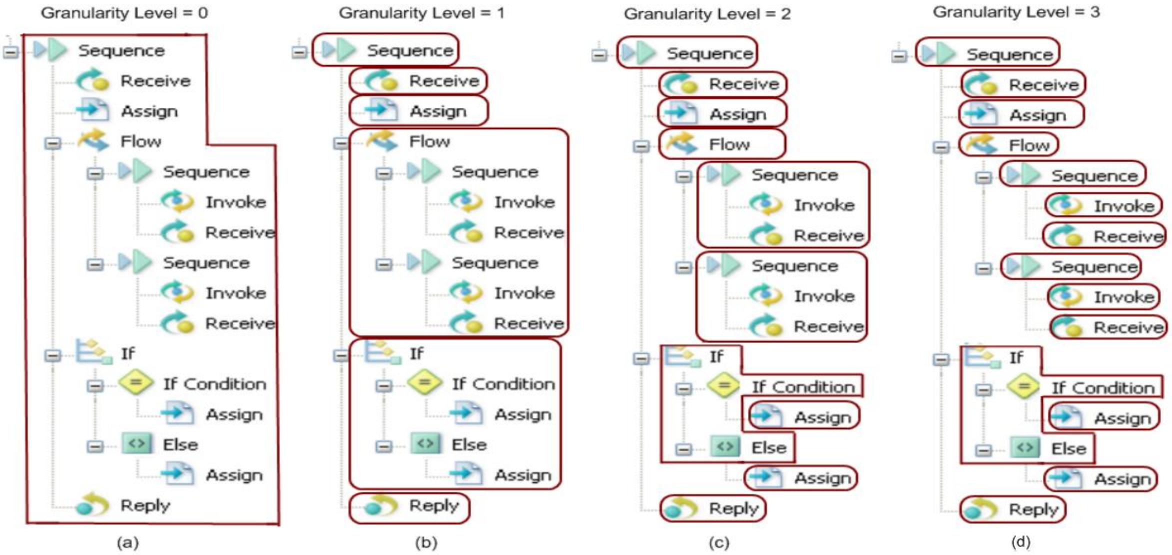


Figure 2: HPD Decentralization of Loan Application Business Process

IV. Adaptable Process Decentralization Framework

A run-time environment is the subject of many changes during the execution of software applications, which may result in violating system thresholds. The adaptability of software with requirements of its execution environment is important in that it helps the run-time environment to refrain from catastrophic events.

Distributed systems are used to resolve the scalability issues of centralized models. However, distribution may result in performance bottleneck due to the communication among system components and continuous changes in a distributed environment. Adaptability of a distributed system, i.e. a decentralized orchestrate engine, with its environment is of high importance to avoid approaching system thresholds and bypassing bottlenecks.

Figure 3 shows the main phases of an Adaptable and Decentralized Workflow Execution Framework (ADWEF) to support the adaptable decentralization of business processes. The central part of the framework is a feedback data repository, which can be initialized by different parties such as a system administrator, a distributed workflow engine, configuration files, monitoring devices and software, etc. Based on the data provided in the feedback repository, a decentralization decision maker may be able to decentralize/re-decentralize a new/running business process. Re-decentralization may occur due to violating system thresholds. Making decision on how to decentralize a business process, the decision maker submits required information to the workflow decomposer which is able to fragment a business process to workflow fragments. Through a process of deployment, the produced fragments are encapsulated into runtime components (i.e., agents) and they will be deployed into the machines dedicated to a distributed workflow engine. Dynamic architectures, which support the execution of dynamic fragments, have to implement collaborative components to reinforce each of the phases specified in the framework.

V. Adaptable Decentralization Decision Maker Unit

This section elaborates an adaptable decentralization decision maker unit. Adaptability may come along with different criteria such as memory usage, bandwidth usage, throughput, etc. Nevertheless, having all of them together is impossible due to confliction of goals. This section also opens discussions on considering run-time circumstances in business process decentralization. Figure 4 also presents a decentralization decision maker unit, which offers a suitable level of decentralization based on receiving two parameters from run-time environment including the number of available machines and available bandwidth. The decision maker unit is implemented using a Fuzzy approach in this paper.

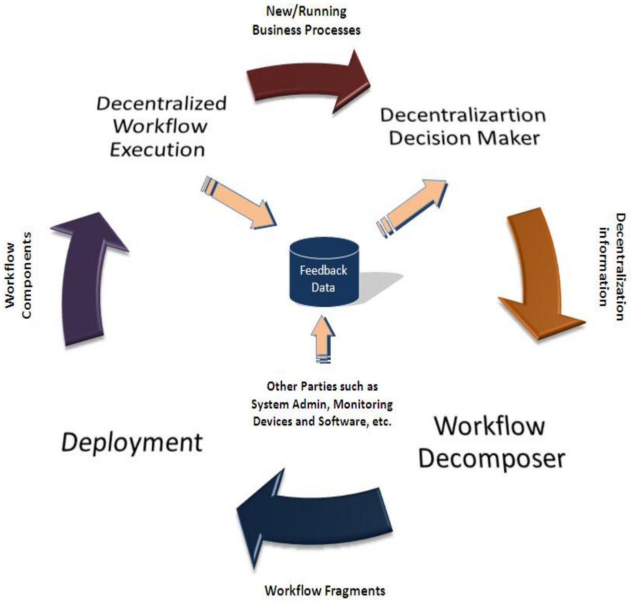


Figure 3: Adaptable and Decentralized Workflow Execution Framework (ADWEF)

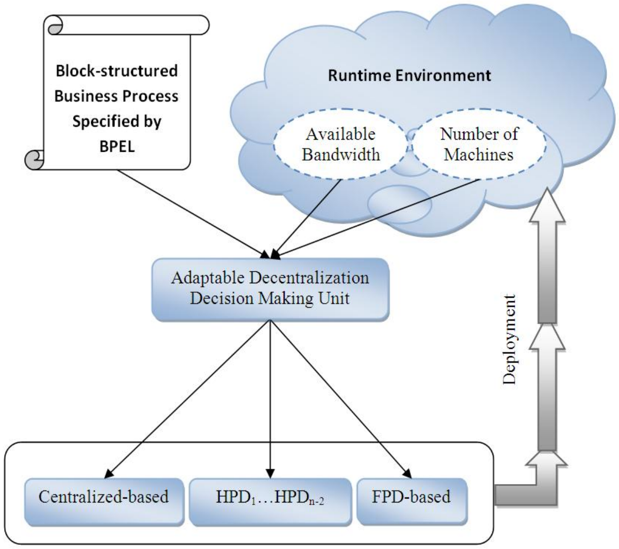


Figure 4: Decentralization Decision Maker

```
name: fuzzyGranularity;
input:
  ps (Process Specification),
  bw (Bandwidth),
  nom (Number of Machines);
output:
  granularityLevel;
begin
  fsa = fragmentSetArray (ps, “HPD”);
  fp = findFragmentProportionality (fsa, nom);
  fpl = findFragmentProportionalityLevel (fp)
  granularityLevel =  findFuzzyGranularity (fsa, fpl, bw);
  return granularityLevel;
end;
```

Figure 5: Adaptable Fuzzy Decentralization Algorithm

VI. Bandwidth Adaptable Decentralization

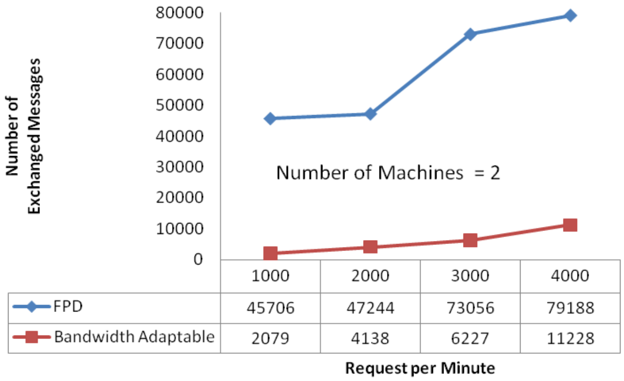
Bandwidth is important in that it is independent from other parameters such as response time and throughput; however, response time and throughput are highly affected by the available bandwidth. A busy communication media may increase the latency of communication among components, which results in increasing response time and decreasing throughput, consequently.

A sample implementation of decision making unit is shown in Figure 5. It shows the *fuzzyGranularity* algorithm as well as its input parameters including process specification, current bandwidth and number of machines. The level of decentralization will be the output parameter. At first, the business process specification is sent to *fragmentSetArray* by identifying the method of fragmentation i.e. HPD; second, fragment proportionality is determined by *findFragmentProportionality*. It determines that the number of produced fragments in which decentralization level is closer to the number of machines. The closer number is called fragment proportionality. Then, the level of a business process tree, which satisfies the fragment proportionality, is returned by the next method *findFragmentProportionalityLevel*.

Finally, by having all the fragments, available bandwidth and fragment proportionality level, granularity level is calculated by the *findFuzzyGranularity* method. This method receives fragment set array (*fsa*), fragment proportionality level (*fpl*) and available bandwidth (*bw*) as input and returns the level of decentralization in process tree as output. At first, bandwidth is segmented dynamically to a set of Segments (S_i) and then; *bw* is fuzzified using a singleton function. For each segmented bandwidth S_i a new rule is created using a singleton function such that $S_i \rightarrow$ *Singleton (fsa[i].fragmentNo())*. The created rules are executed using a rule engine. Output is defuzzified later and finally a crisp value is calculated. The crisp value determines a suitable level of process tree, which is the final result of *findFuzzyGranularity* method.

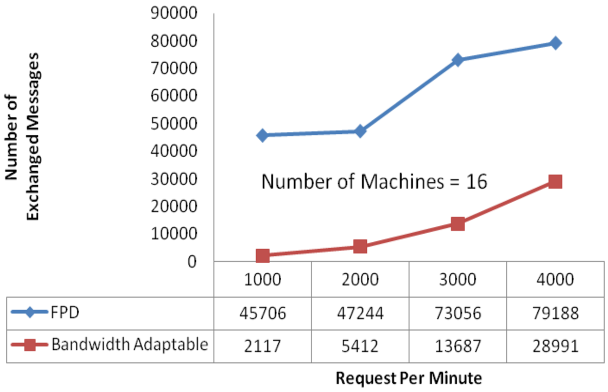
VII. Experimental Setup and Evaluation

The focus of this section is evaluating the behavior of the bandwidth-adaptable fuzzy decentralization decision making unit. The algorithm is expected to adapt decentralization of processes with current available bandwidth. It considers the fragment proportionality of decentralization with number of machines as well.



	1000	2000	3000	4000
FPD	45706	47244	73056	79188
Bandwidth Adaptable	2079	4138	6227	11228

Figure 6: Comparing exchanged messages by fragments of FPD and the presented algorithm using two machines.



	1000	2000	3000	4000
FPD	45706	47244	73056	79188
Bandwidth Adaptable	2117	5412	13687	28991

Figure 7: Comparing exchanged messages by fragments of FPD and presented algorithm using sixteen machines.

In order to implement the experiments, WADE/JADE [23-26] platform was selected and installed on a network with sixteen machines. The created fragments using the HPD method were encapsulated in WADE/JADE agents and deployed to network machines. Accordingly, the experiment was repeated for two and sixteen machines. A client sent requests with specific rates of 1000, 2000, 3000 and 4000 per minute. During the experiments, sniffer software monitored the number of messages exchanged among the fragments. Receiving a request from a client, a number (between 0-100) was generated with exponential distribution, which was the simulation of available bandwidth. The available bandwidth along with the fragment proportionality parameter was sent to a fuzzy algorithm to recommend a suitable level of decentralization. The same experiments were repeated without the presence of bandwidth adaptable algorithm, which was analogous to applying the FPD method.

Both Figure 6 and Figure 7 show the results of the experiments. The fully process decentralization method (FPD) was neutral to the number of machines and bandwidth fluctuations; thus, a constant number of messages were passed among the agents. In contrast with the FPD, the bandwidth adaptable algorithm, decentralized the loan process based on the generated bandwidth and number of machines. As shown in Figure 2, different versions of the loan process were created at run-time. The adaptable algorithm reduced the number of exchanged messages in both cases due to considering the adaptability of fragments with number of machines and available bandwidth. In the case that enough bandwidth was available, the algorithm considered only the adaptability with number of machines. When the bandwidth was not wide enough, the algorithm shrunk the fragments and created more centralized fragments.

VIII. CONCLUSION

An adaptable business process decentralization framework was introduced as a solution for decentralizing large scale business processes. The main problem was that current decentralization methods did not consider the adaptability of business process decentralization with run-time environment circumstances such as available bandwidth in this paper.

By decentralizing a business process based on the number of machines, the extra communication cost of inter-fragment communication is omitted. If there is only one machine to execute processes, there is no need to create several fragments running on one machine. In other words, a multi-thread process would be more effective. In the case of several machines, the proportionality of produced fragments with number of machines dedicated to a distributed engine reduces the number of fragments and consequently communication among them. On the other hand, the available bandwidth of a network may directly affect the response time and throughput of workflows. Imagine a busy communication media in a distributed workflow engine. Under such circumstances, creating a fully fragmented process may result in a huge amount of exchanging messages among the fragments and then system approaches/violates its thresholds, consequently. Obviously, the less number of fragments provides better outcomes in this case. As a matter of fact, the process is shrunk to refrain from violating the thresholds. After ameliorating the run-time conditions, current processes and/or new processes, may be expanded/re-expanded by creating more fragments.

Indeed, in this paper, an adaptable process decentralization framework was introduced to create fragments proportional to both the number of machines and current available bandwidth. The evaluation of bandwidth adaptable algorithm showed that this algorithm was able to execute a process with considerably less number of exchanged messages compared to the FPD method. The fragments of the FPD exchanged ten to twenty times more messages compared to the bandwidth adaptable algorithm.

It is worth mentioning that the adaptable decentralization decision making unit opens more discussions on finding more adaptability metrics and more intelligent algorithms to achieve better decentralization outcomes. Considering the network capacity, accumulative bandwidth, the relation of workflow activities, etc can be instances of such adaptability metrics. Furthermore, algorithms are required to decentralize graph-structured business processes and mapping them to run-time circumstances. Currently, our main focus is on developing a Fuzzy algorithm to integrate the HPD and HIPD methods to provide a more adaptable decentralization approach.

ACKNOWLEDGMENT

Faramarz Safi Esfahani is also a PhD candidate in University of Putra Malaysia, 43400, Serdang, Selangor, Malaysia.

REFERENCES

[1] Guoli Li, Vinod Muthusamy, and Hans-Arno Jacobsen, “A Distributed Service Oriented Architecture for Business Process Execution,” ACM Transactions on the Web (TWEB), vol. 4, no. 1, article 2, 2010.

[2] Paolo Bruni, Marcos Henrique Simoes Caurim, Alexander Koerner, Christine Law, et al., Powering SOA with IBM Data Servers, ISBN. 738494542, IBM, 2006, p. 754.

[3] “Workflow management coalition: process definition interchange,”, http://www.wfmc.org, 2011.

[4] Mirkov Viroli, Enrico Denti and Alessandron Ricci, “Engineering a BPEL orchestration engine as a multi-agent system,” Journal of Science of Computer Programming, vol 66, issue 3, 2007, pp. 226-245 2007.

[5] Roberto Silveira Silva Filho, Jacques Wainer and Edmundo Roberto Mauro Madeira “A fully distributed architecture for large scale workflow enactment,” International Journal of Cooperative Information Systems, vol. 12, no. 4, 2003, pp. 411-440.

[6] Giancarlo Fortino, Alfredo Garro, Wilma Russo, “Distributed Workflow Enactment: an Agent-based Framework,” Proc. WOA2006, 2006, pp. 110-117.

[7] Li Guo, Dave Robertson, Yun-Heh Chen-Burger, “A Novel Approach for Enacting the Distributed Business Workflows Using BPEL4WS on the Multi-Agent Platform,” Web Intelligence and Agent Systems, 2005, pp. 657-664.

[8] Daniel Wutke, Daniel Martin and Frank Leymann, “Model and Infrastructure for Decentralized Workflow,” Proc. ACM/SAC 2009, ACM, 2008, pp. 90-94.

[9] Gustavo Alonso, Fabio Casati, kuno Harumi and Machiraju Vijay, “Exotica/FMQM: A Persistent Message-Based Architecture for Distributed

Workflow Management," Proc. IFIP WG 8.1 Workgroup Conference on Information Systems Development for Decentralized Organizations (ISD095), 1995, pp. 1-18.

[10] Vinod Muthusamy, Hans-Arno Jacobsen, Tony Chau and Allen Chan, "SLA-driven business process management in SOA," Proc. CASCON, Toronto, Canada, 2009, pp. 86-100.

[11] Ting Cai, Peter A Gloorand and Saurab Nog, "DartFlow: A workflow management system on the web using transportable agents," Technical Report PCS-TR96-283, Dartmouth College, Hanover, NH, 1996, http://www.cs.dartmouth.edu/~jcrespo/cms_file/SYS_techReport/156/TR96-283.pdf.

[12] Wei Tan and Yushun Fan, "Dynamic workflow model fragmentation for distributed execution," Comput. Ind., vol. 58, no. 5, 2007, pp. 381-391; DOI http://dx.doi.org/10.1016/j.compind.2006.07.004.

[13] Vijayalakshmi Atluri, Soon Ae Chun, Ravi Mukkamala and Pietro Mazzolen, "A decentralized execution model for inter-organizational workflows," Distrib. Parallel Databases, vol. 22, no. 1, 2007, pp. 55-83; DOI http://dx.doi.org/10.1007/s10619-007-7012-1.

[14] Faramarz Safi Esfahani, Masrah Azrifah Azmi Murad, Md. Nasir Sulaiman and Nur Izura Udzir," Adaptable Distributed Service Oriented Architecture," Elsevier Journal of Systems and Software (JSS), in press.

[15] Luciano Baresi, Elisabetta Di Nitto and Carlo Ghezzi, "Toward open-world software: Issue and challenges," Computer, vol. 39, no. 10, 2006, pp. 36-43.

[16] Yiwei Gong, Marijn Janssen, Sietse Overbeek and Arre Zuurmond, "Enabling flexible processes by ECA orchestration architecture," Proc. ICEGOV, 2009, pp. 19-26.

[17] Faramarz Safi Esfahani, Masrah Azrifah Azmi Murad, Md. Nasir Sulaiman and Nur Izura Udzir, "Using Process Mining To Business Process Distribution," Proc. SAC2009, ACM, 2009, pp. 1876-1881.

[18] Faramarz Safi Esfahani, Masrah Azrifah Azmi Murad, Md. Nasir Sulaiman and Nur Izura Udzir, "An Intelligent Business Process Distribution Approach," Journal of Theoretical and Applied Information Technology, vol. 4, 2008, pp. 1236-1245.

[19] Faramarz Safi Esfahani, Masrah Azrifah Azmi Murad, Md. Nasir Sulaiman and Nur Izura Udzir, "SLA-Driven Business Process Distribution," Proc. IARIA/eKnow2009, IEEE, 2009, pp.14-21.

[20] Faramarz Safi Esfahani, Masrah Azrifah Azmi Murad, Md. Nasir Sulaiman and Nur Izura Udzir, "A Case Study of the Intelligent Process Decentralization Method," Proc. WCECS, IAENG, 2009, pp 269-274.

[21] Rania Khalaf and Frank Leymann, "E Role-based Decomposition of Businesses using BPEL," Proc. IEEE International Conference on Web Services (ICWS'06), 2006, pp. 770-780.

[22] Mangala Gowri Nanda, Satish Chandra and Vivek Sarkar, "Decentralizing execution of composite web services," ACM SIGPLAN Notices, vol. 39, no. 10, 2004, pp. 170-187.

[23] Fabio Bellifemine, Agostino Poggi and Giovanni Rimassa, "JADE–A FIPA-compliant agent framework," Proc. PAAM99, 1999, pp. 97-108.

[24] Giovanni Caire, Danilo Gotta and Massimo Banzi, "WADE: a software platform to develop mission critical applications exploiting agents and workflows," Proc. 7th international joint conference on Autonomous agents and multiagent systems: industrial track, Richland, SC, 2008, pp. 29-36.

[25] Krzysztof Chmiel, Maciej Gawinecki, Pawel Kaczmarek, Michal Szymczak, et al., "Efficiency of JADE agent platform," Journal of Scientific Programming, vol. 13, no. Number 2/2005, 2005, pp. 159-172.

[26] Bellifemine Fabio, Caire Giovanni and Greenwood Dominic, Developing Multi-Agent Systems with JADE, WILEY, 2007.

Keyphrase Extraction by Synonym Analysis of *n*-grams for E-Journals Categorisation

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

***Abstract*—Automatic keyword or keyphrase extraction is concerned with assigning keyphrases to documents based on words from within the document. Previous studies have shown that in a significant number of cases author-supplied keywords are not appropriate for the document to which they are attached. This can either be because they represent what the author *believes* the paper is about not what it actually is, or because they include keyphrases which are more classificatory than explanatory e.g., "University of Poppleton" instead of "Knowledge Discovery in Databases". Thus, there is a need for a system that can generate appropriate and diverse range of keyphrases that reflect the document. This paper proposes a solution that examines the synonyms of words and phrases in the document to find the underlying themes, and presents these as appropriate keyphrases. The primary method explores taking n-grams of the source document phrases, and examining the synonyms of these, while the secondary considers grouping outputs by their synonyms. The experiments undertaken show the primary method produces good results and that the secondary method produces both good results and potential for future work.**

Keywords- Automatic tagging, Document classification, Keyphrases, Keyword extraction, Single document, Synonyms, Thesaurus

I. INTRODUCTION

Keywords are words used to identify a topic, theme, or subject of a document, or to classify a document. They are used by authors of academic papers (such as all papers about "metaphor" or "leadership"), by libraries to allow people to locate books (such as all books on "Stalin" or "romance"), and other similar uses. The keywords for a document indicate the major areas of interest within it.

A keyphrase is a short phrase of, perhaps, one to five words, which fulfils a similar purpose, but with broader scope for encapsulating a concept. While this may be considered the authors' contention, it is inferred that a short phrase of a few linked words contains more meaning than a single word alone, e.g., the phrase "natural language processing" is more useful than just the word "language".

Frank et al. [1] discuss two different ways of approaching the problem of linking keyphrases to a document. The first, keyphrase assignment, is to assume a set and given list of keyphrases or categories which can be assigned to the document. The computational problem for this approach is then to determine a mapping between documents and categories using already classified documents as learning aids. The second approach, keyphrase extraction, assumes there is no restricted list and instead attempts to use any phrase from the document (or ones constructed via a reference document) to serve as the keyphrases.

Previous research [2][3] has shown that for any given group of documents with keyphrases, there is a small number which are frequently used (examples include "shopping" or "politics" [3]) and a large number with low frequency (examples include "insomnia due to quail wailing" or "streetball china" [3]). The latter set is too idiosyncratic for widespread use; generally, even reuse by the same author is unlikely. Therefore, part of the issue of both keyphrase assignment and extraction is locating the small number of useful keyphrases to apply to the documents.

This project is concerned with keyphrase extraction and, as such, this paper covers the background research into keyword/keyphrase generation, outlines a proposed solution to the problem, and compares the performance to manually assigned keyphrases. The main aim is to take an arbitrary document (in isolation from a corpus) and analyse the synonyms of word-level *n*-grams to extract automatically a set of useful and valid keywords, which reflect the themes of that document. The words of the document are analysed as a series of *n*-grams, which are compared to entries in a thesaurus to find their synonyms and these are ranked by frequency to determine the candidate keywords. The secondary aim is to look at a method of grouping the theme outputs into clusters, so that the results did not just show the most common theme swamping out any others.

The rest of the paper comprises the background and state-of-the-art (Section II), the implementation and results gained (Section III), a discussion (Section IV), and conclusions and suggestions for future work (Section V).

II. BACKGROUND

The background research into automatic keyword generation has shown that existing work in these areas focuses on either cross analysing a corpus of multiple documents for conclusions or extrapolating training data from manual summaries for test documents. While manual summaries generally require multiple documents to train upon they do not need to compare each component of the corpus to all other components. Instead, they try to

ISBN: 978-1-61208-121-2

extrapolate the patterns between the pairs of documents and manual summaries in the training set.

A. *Single Documents*

Single document approaches make use of manual summaries or keyphrases to achieve their results. Tuning via manual summaries attempts to replicate the process by which a human can identify the themes of a document and reduce the text down to a summary/selection of keyphrases. The general approach taken involves a collection of documents (with associated human summaries) and a given method is applied to draw relationships between the document and the summary. From this, new documents (generally a test corpus that also contains human summaries) are subject to the derived relationships to see if the summaries produced by the system are useful and usable.

For creating summaries, Goldstein et al. [4] set out a system based upon assessing every sentence of the document and calculating a ranking for its inclusion in a summary. The authors made use of corpora of documents for which assessor-ranked summary sentences already existed, and attempted to train the system to produce similar or identical sentences.

A different approach was taken by the *Stochastic Keyword Generator* [5], a proposed system for classifying help desk problems with short summaries. Submitted e-mails varied in their description of the problem and often contained duplicated or redundant data. Therefore, the authors created a system that would attempt to create a summary similar to those manually created by the help desk staff: concise, precise, consistent, and with uniform expressions. Their system uses a corpus of e-mails with manual summaries and ranks source words for inclusion based on the probability that they will occur based on the probability from its training data.

For producing keyphrases, Barker and Cornacchia [6] propose a system that takes into account not only the frequency of a "noun phrase" but also the head noun. For example, tracking "the Canadian Space Agency" should also track counts of "the Space Agency" or "the Agency". Wermter and Hahn [7] examine a method of ranking candidate keyphrases using the limited paradigmatic modifiability (LPM) of each phrase as a guide to locating phrases with low frequency but high interest to the document.

B. *Multiple Documents*

Multiple document approaches take a corpus and attempt to analyse relationships between the component elements to create methods for dealing with unseen elements. Most of these approaches are based on examining parts of an individual document in the corpus and then examining how that differs across the other documents.

"*TagAssist*" [2] makes use of a continually updated corpus of blog posts (supplied by [3]) and author-supplied tags to suggest tags for new blog posts. The system compares the author's tags and content of blog posts to work out the relationships that prompt the former to be chosen to represent the latter. Their baseline system worked on a simple frequency count for determining output. Evaluated by ten human judges (unaware of which system produced each tags), the results showed that the original tags were the most appropriate (48.85%) with *TagAssist* coming in second (42.10%), and the baseline system last (30.05%).

The *C-Value* and *NC-Value* [8] are presented as methods for ranking "term words" taking into account phrase length and frequency of its occurrence as a sub-string of another phrase. *TRUCKS* [9] extends the *NC-Value* work, combining it with [10], to use contextual information surrounding the text to improve further the weightings used in the *NC-Value*.

Extra data may be used to gain more information on the relationships between the components, often gained from reference documents. Joshi and Motwani [11] make use of a thesaurus to obtain extra meaning from keywords so their program, "*TermsNet*", can observe keywords in their original context in attempt to link said keywords in "non-obvious ways". Scott and Matwin [12] use the *WordNet* lexical database [13] to find the hyponyms and feed this information to the Ripper machine learning system. Wei et al. [14] demonstrate such a system that uses *WordNet* to generate keywords for song lyrics. Their approach clusters the words of a song using *WordNet's* data to link words across the song. Keywords are then found at the centres of these links.

III. Implementation and Results

The basis of this work is the examination of a document with reference to its synonyms and therefore the main bulk of the coding of the system related to this and the associated thesaurus file. The input thesaurus corpus for analysis was Roget's "*Thesaurus of English Words and Phrases*" [15] and was chosen due to availability and because initial prototyping had shown that WordNet [13] (normally used within the discipline) performed less well for this application (see also Section V).

The system was tested on a number of papers taken from a collection of online e-journals, Academics Conferences International (ACI) [16]. There were five e-journals in this collection, each on a different topic and they were analysed separately. The topics were *Business Research Methods* (EJBRM), *E-Government* (EJEG), *E-Learning* (EJEL), *Information Systems Evaluation* (EJISE), and *Knowledge Management* (EJKM).

For each of these e-journals the authors supply keywords/phrases. The baseline evaluation of this work is to compare the keyphrases supplied by the author with those identified by the system under consideration. A match is assumed if one author's keyphrase matches a system-supplied keyphrase using a naïve text-matching method. This method would match the word "know" with both the words "know" and "knowledge".

Table I shows the baseline results for the study, which were established by running the system with only unigrams and no clustering, and outputting only the most common keyphrase for each paper.

For each of the methods described below the thesaurus was loaded into the program and stored as a list of linked pairs of data, consisting of a unique Key (base word in the thesaurus) and an associated Value (its synonyms). The keys ranged from unigram word entries up to 7-gram phrases.

TABLE I. BASE LINE RESULTS

Journal	Papers	Matched	Percentage
EJBRM	72	2	2.78%
EJEG	101	3	2.97%
EJEL	112	16	14.29%
EJISE	91	6	6.59%
EJKM	110	15	13.64%
Average			8.47%

The project was split into two methods: the *n*-gram study and the clustering study. The following sections outline these approaches and the results from each.

A. *The n-gram study*

For the *n*-gram study, the words from the source document were split into a number of *n*-gram lists, from unigrams up to 7-grams. For all of the lists the entries overlapped so that all combinations of words from the text were included. E.g., if the source text were "The quick fox jumped" then the bigrams would be "The quick", "quick fox", and "fox jumped" and the trigrams would be "The quick fox", and "quick fox jumped". For each document, the results of each of the *n*-grams were combined and considered together to determine the overall output.

Each time the *n*-gram appeared in the source text, its frequency in its word list was increased by n. The unigrams were then stemmed (to remove plurals, derivations, etc.) using the Porter Stemming Algorithm [17], and added to the list with combined frequencies from each of the unigrams that reduced to that stem. The resultant corpus of *n*-grams and stems was then compared to the entries in the thesaurus:

- For each word (Key) in the thesaurus, compare the *n*-gram to the associated synonyms (Value).
- For each synonym that matches, add the word (Key) to a list, and increase its frequency value by the value of the *n*-gram.
- Sort the list by frequency and output the top *r* ranked items (in this study, *r* was chosen to be 5).

The *n*-gram results showed a reasonable improvement (52%) over the baseline, as can be seen in Table II. The increase measures the performance compared to the results from Table I.

TABLE II. RESULTS OF *N*-GRAM STUDY

Journal	Papers	Matched	Percentage	Increase
EJBRM	72	25	34.72%	31.94%
EJEG	101	71	70.30%	67.33%
EJEL	112	72	64.29%	50.00%
EJISE	91	39	42.56%	35.97%
EJKM	110	94	85.45%	71.84%
Average			60.69%	52.22%

B. *The clustering study*

The clustering algorithm attempts to extend the *n*-gram algorithm to group the keyphrases into "clusters" by finding the keyphrases that are of a similar theme and returning a single keyphrase for that group. For example, the word "recovery" can mean either "acquisition" or "taking" [15]. The base system therefore could return multiple versions of the same concept as keyphrases. By clustering the results, the attempt was to prevent a single, "popular", concept dominating and allow the other themes to be represented. The method for this was:

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

The clustering results show only a small improvement over the *n*-gram alone, as can be seen in Table III. The increase measures the performance compared to the results from Table I.

TABLE III. RESULTS OF CLUSTERING STUDY

Journal	Papers	Matched	Percentage	Increase
EJBRM	72	31	43.06%	40.28%
EJEG	101	73	72.28%	69.31%
EJEL	112	77	68.75%	54.46%
EJISE	91	46	50.55%	43.96%
EJKM	110	94	85.45%	71.81%
Average			64.72%	56.25%

IV. DISCUSSION

The results show that while using *n*-grams on their own produce an increase in the matched output, extending the algorithm to include clustering produced a small improvement. However, when the clustering algorithm was run on a system using only unigrams, it performed just as well (identically in fact) as with a greater number of *n*-grams. This suggests that the clustering algorithm is of more use to the project aims than the initial *n*-gram work, and provides an interesting area to expand upon in future work.

In addition to this, some of the keywords submitted by the authors may be tags instead and these display meta-data that can often be irrelevant to the understanding of the document. An example seen in the corpus was the keyword "University of Birmingham" because the author of that paper worked there. This is valid as a tag but as a keyword, it does not indicate a topic or a theme to which the document holds (other than in a rare case where the paper is about the University of Birmingham). This would therefore lower the chances of keyphrases being matched as the comparison data is filled with `noise'.

The synonyms are currently analysed context-free, and thus for a word with multiple meanings (e.g., "recovery" can mean "acquisition", "improvement", or "restoration" [15]) every occurrence of that word is treated the same. This means that a document equally about "improvement" and "restoration" could end up with the theme of "recovery" which (while a correct assumption) may not give the right meaning.

V. CONCLUSION AND FURTHER WORK

The approach to synonym analysis developed in this paper showed good results for the test corpora used and potential for future study. Further study is required to compare the system to ones developed in similar areas, but this should provide a solid framework for taking the project forward.

The results also show that the use of *n*-grams, while useful on its own, has no effect upon the percentage match for the system. This does not, however, mean that the keywords produced may not be more useful to the user, as they could be different enough not to match the success criteria but still relevant.

The results themselves were evaluated against the keywords submitted by the authors of the papers. *TagAssist* [2] showed that in 54.15% of cases, author keywords were judged as being inappropriate for the work with which they were associated. Therefore, when interpreting the results (which averaged around 60% matches) it should be remembered that they are produced by matching the output against the author keywords, which may be less than perfect for the task. A new method of evaluating the results is, therefore, needed.

Another area of further work for this project is conducting more experiments to determine why the *WordNet* thesaurus performed worse in initial trials, and whether it is better at certain subject classes of documents (e.g., a medical corpus vs. a computer science corpus).

ACKNOWLEDGMENT

The authors would like to thank the School of Systems Engineering for the studentship, which enabled this project, and the contributions from the reviewers to this paper.

REFERENCES

[1] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. 1999. "Domain-Specific Keyphrase Extraction", Proceedings 16th International Joint Conference on Artificial Intelligence, pp. 668–673. San Francisco, CA Morgan Kaufmann Publishers.

[2] S.C. Sood, S.H. Owsley, K.J. Hammond, and L. Birnbaum. 2007. "TagAssist: Automatic Tag Suggestion for Blog Posts". Northwestern University. Evanston, IL, USA. http://www.icwsm.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf [Last accessed: 13 December 2010]

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

[4] J. Goldstein, M. Kantrowtiz, V. Mittal, and J. Carbonell. 1999. "Summarising Text Documents: Sentence Selection and Evaluation Metrics", ACM, pp. 121–128. Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA.

[5] C. Li, J. Wen, and H. Li. 2003. "Text Classification Using Stochastic Keyword Generation", Twentieth International Conference on Machine Learning (ICML), pp 464–471. Washington DC. https://www.aaai.org/Papers/ICML/2003/ICML03-062.pdf [Last accessed: 13 December 2010]

[6] K. Barker and N. Cornacchia. 2000. "Using Noun Phrase Heads to Extract Document Keyphrases", AI '00: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence. pp. 40–52). London.

[7] J. Wermter and U. Hahn. 2005. "Paradigmatic Modifiability Statistics for the Extraction of Complex Multi- Word Terms". Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) pp. 843–850. Vancouver Association for Computational Linguistics.

[8] K. Frantziy, S. Ananiadou, and H. Mimaz. 2000. "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method", International Journal on Digital Libraries , 3 (2), pp. 117-132.

[9] D. Maynard and S. Ananiadou. 2000. "TRUCKS: a model for automatic multi-word term recognition". Journal of Natural Language Processing, 8 (1), pp. 101-125.

[10] D. Maynard and S. Ananiadou. 1999. "Term extraction using a similarity-based approach". Recent Advances in Computational Terminology, pp. 261–278.

[11] A. Joshi and R. Motwani. 2006. "Keyword Generation for Search Engine Advertising", IEEE International Conference on Data Mining, pp. 490–496.

[12] S. Scott and S. Matwin. 1998. "Text Classification Using WordNet Hypernyms", Proceedings of the Association for Computational Linguistics, pp. 38–44.

[13] G.A. Miller, C. Fellbaum, R. Tengi, P. Wakefield, and H. Langone. 2005. "WordNet". Princeton University. http://WordNet.princeton.edu [Last accessed: 13 December 2010]

[14] B. Wei, C. Zhang, and M. Ogihara. 2007. "Keyword Generation for Lyrics", Austrian Computer Society (OCG). Comp. Sci. Dept., U. Rochester, USA. http://ismir2007.ismir.net/proceedings/ISMIR2007_p121_wei.pdf [Last accessed: 13 December 2010]

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

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

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