BUSTECH 2011

The First International Conference on Business Intelligence and Technology

ISBN: 978-1-61208-160-1

September 25-30, 2011
Rome, Italy

BUSTECH 2011 Editors

Maribel Yasmina Santos, University of Minho, Portugal

Vagan Terziyan, University of Jyvaskyla, Finland

BUSTECH 2011

Foreword

The First International Conference on Business Intelligence and Technology [BUSTECH 2011], held between September 25 and 30, 2011 in Rome, Italy, initiated a series of events covering topics related to business process management and intelligence, integration and interoperability of different approaches, technology-oriented business solutions and specific features to be considered in business/technology development.

The term Business Intelligence (BI) covers a large spectrum of applications and technologies used to collect, store, interpret and decide on the information about company operations with the aim of helping corporate entities with a comprehensive status and knowledge on their business. BI is integrating with Warehouses (DWs), on-line analytic (OLAPS), corporate performance management (CPM), business process management (BPM), and other technology-oriented business solutions. Web technologies, semantics and ontology mechanisms are now used to mine, integrate and interpret distribute corporate data, either real-time or intermittent, by filtering noisy data, interpreting business data in context, enforcing trust and security in handling corporate data, and providing access to data from anywhere, at anytime, and via any media. The complexity, volume and intrinsic semantic of data needed for conducting business require a tailored IT infrastructure and advanced methodologies and technologies for timely building competitiveness by intelligent business decisions. With the large spectrum of emerging technologies, such as cloud computing, sensors environments, and mobility there is a need for specialized supporting tools and business/technology decisions to optimize the business process and business performance.

We take here the opportunity to warmly thank all the members of the BUSTECH 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 BUSTECH 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 BUSTECH 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BUSTECH 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 area of business intelligence and technology.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Rome, Italy.

BUSTECH 2011 Chairs:

BUSTECH 2011

Committee

BUSTECH Advisory Chairs

Lars Taxén, Linköping University, Sweden
Dieter Fensel, STI Innsbruck / University of Innsbruck, Austria
Malgorzata Pankowska, University of Economics – Katowice, Poland
Hannes Werthner, Vienna University of Technology, Austria

BUSTECH 2011 Research/Industry Chair

Tyrone W. A. Grandison, IBM Research - Hawthorne, USA

BUSTECH 2011 Special Area Chairs

Semantic/Ontology
Oscar Ferrandez Escamez, University of Utah, USA

Business-driven IT
Maribel Yasmina Santos, University of Minho, Portugal

Security
Pedro Soria-Rodriguez, Atos Origin - Madrid, Spain

Business continuity
Sokratis K. Katsikas, University of Piraeus, Greece

Services
Michael Parkin, European Research Institute in Service Science (ERISS), The Netherlands

BUSTECH 2011 Technical Program Committee

Witold Abramowicz, The Poznan University of Economics, Poland
Hassan Ait-Kaci, IBM, Canada
Antonia Albani, University of St. Gallen, Switzerland
Eric Andonoff, IRIT / Université Toulouse 1, France
Renata Araujo, Federal University of the Rio de Janeiro State (UNIRIO), Brazil
Colin Atkinson, University of Mannheim, Germany
Anteneh Ayanso, Brock University - St. Catharines, Canada
Fernanda Baiao, Federal University of the State of Rio de Janeiro, Brazil
Joseph Barjis, Delft University of Technology, The Netherlands
Kawtar Benghazi, Universidad de Granada, España
Salima Berbernou, Universty Paris Descartes, France
Stefan Biffl, TU Wien, Austria

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

SBVR Based Representation of SPARQL Queries and SWRL Rules for Analyzing Semantic Relations

Algirdas Sukys, Lina Nemuraite, Bronius Paradauskas, Edvinas Sinkevicius
Kaunas University of Technology
Kaunas, Lithuania
sukys.algirdas@gmail.com, lina.nemuraite@ktu.lt, bronius.paradauskas@ktu.lt, edvis.s@gmail.com

***Abstract*—Analyzing semantic relations is a cumbersome task because these relations often are distributed over different information sources and hidden in existing relational database structures. Even with Semantic Web ontologies describing semantic relations we need to explore them on the deep technological level that is not friendly for business users. Semantics of Business Vocabulary and Business Rules (SBVR) gives a possibility of representing OWL 2 ontologies, SWRL rules and SPARQL queries using concepts and semantic formulations expressed in SBVR Structured English language understandable for business users. We suggest formulating derivation rules and queries for analyzing semantic relations as SBVR rules and questions, and transforming them into SWRL and SPARQL.**

Keywords-semantic relations; SBVR; SPARQL; OWL 2; SWRL.

I. INTRODUCTION

More and more business information becomes available in form of ontologies that are accessible to users by dedicated query languages as SPARQL [24][32]. However, users prefer querying using sentences in natural language [10]. Query languages are limited, therefore complex parsing and validating means are needed for ensuring flexibility of queries that should allow using synonyms and synonymous forms, and asking in various ways. Ontology reasoners could help to query using expressions that are not directly defined in ontology but could be derived on the base of existing formulations.

Semantics of Business Vocabulary and Business Rules (SBVR) is the OMG metamodel that defines the vocabulary and rules for describing the business semantics – business concepts, business facts, and business rules using SBVR Structured English (SSE) or other Controlled Natural Language [21]. The SSE language cannot represent all possible constructions of natural English language, however, it is understandable to business and information system experts, and is interpretable by computers. SBVR directly models business meaning and can support many controlled languages; consequently, it can be seen as an interface between real natural languages and business software systems [11].

SBVR presents a high-powered metamodel but a lot of work should be done for introducing SBVR based business semantics management into the enterprise architecture. SBVR is capable to represent the human understandable semantics beyond Web Ontology Language (OWL 2), Semantic Web Rules (SWRL), OWL for Web Services (OWLS), Web Service Modeling Ontology (WSMO), and other ontological languages that were acknowledged as most expressive semantic representations before SBVR. SBVR directly models meaning and offers possibility to relate business semantics to software models and implementations; however, such a relation could be made through multiple transformations between various languages and architectural layers.

Despite the interest from the researchers and business side, until now only part of SBVR was used. SBVR questions provide a capability of querying business models and their implementations but they attained only little attention in the SBVR specification and related research. Kriechhammer in 2006 has noticed about possibilities of SBVR questions [12] for business people to query systems for business modeling without the support of programmers. To our knowledge, no further research in that direction was done. In [29], we applied the idea and presented the initial methodology for transforming SBVR questions into SPARQL.

In the current paper, we analyze further possibilities of querying business systems by supplementing previously described querying capabilities with the use of SWRL derivation rules together with OWL 2. We concentrate on querying some types of semantic relations: reflexive relations causing hierarchical links between individual concepts, and n-ary relations that cannot be directly represented in ontology and, consequently, are problematic for querying with SPARQL if we want to obtain SPARQL queries from SBVR questions acceptable for human.

The rest of the paper is organized as follows. Section 2 presents related works. Section 3 presents SBVR questions, corresponding SWRL rules and SPARQL queries for hierarchical relations, Section 4 – for relation properties allowing defining n-ary relations in OWL 2. Section 5 draws conclusions and outlines the future work.

II. RELATED WORK

SBVR metamodel and XMI schema may be used for developing software tools for managing business vocabularies as well as for automating development of software for managing business on the base of business semantics, i.e., in the way different of previously existed approaches, e.g., [18][22]. SBVR business vocabularies are

 ISBN: 978-1-61208-160-1

transformable into UML&OCL [3][17] and vice versa [2]; BPMN [1], RDB schemas [16], OWL [5], Web services [6], [9] etc. Besides automating the development of software models and code [14][15], SBVR Structured English may serve for creating semantic specifications of legacy information resources, integrating these resources, implementing contextualized and multilingual information systems, etc. The power of SBVR is disclosed by the fact that SBVR specification itself is formally written in SSE [13].

For practical usage, SBVR suffers from various limitations and the lack of tools as editors, validation mechanisms and inference engines. It would be desirable having the larger collection of data types and patterns for constructs needed for expressing arithmetic operations, data and time, past and future events, and similar. Spreeuwenberg and Anderson notice more deficiencies of SBVR: lack of inference; lack of references (rules should be stated in one sentence); necessity to introduce concepts before referencing to them; impossibility to express directives, etc. [26][27]. Individual researchers and their groups have proposed various SBVR extensions, but these extensions have not resulted in a new version of SBVR specification yet.

Realizing the idea of querying business in SBVR requires a creation of a whole infrastructure including tools for authoring SBVR business vocabularies and rules, transforming them into various software models and code, including OWL, SQL, Web Services, business process execution languages, and so on. Several EU projects are devoted for this purpose: OPAALS (2006–2010, generating Web services and data models from SBVR specifications [16][25]), ONTORULE (2009–2012, aiming at integrating knowledge and technologies needed for extracting ontologies and business rules from various documents, including natural language texts; managing them and implementing in software systems). The commercial tool suite for Business Semantics Management Collibra presents capabilities for authoring SBVR vocabularies and rules, generate ontologies and various models of information systems.

In our previous work, we focused on generating UML&OCL models from SBVR specifications. We proposed the methodology for specifying information system requirements on the base of SBVR business vocabularies and business rules related with business process models, and implemented the prototype of tool VeTIS supporting that methodology [3][17]. VeTIS tool recognizes SBVR concepts (object types, roles, fact types, fact type roles, individual concepts) and business rules (various kinds of semantic formulations) that make foundations for conceptualizing business and correspond to knowledge and metaknowledge level [19]. For representing SBVR questions, the extension of the VeTIS tool was needed for including business facts – instances of fact types (ground facts) and propositions (instances of complex semantic formulations based on several fact types) comprising the bottom knowledge level – fact level [20]). Representing business facts is crucial for implementing the semantic enterprise where you are able not only tracing business requirements to their implementations in software, but also accessing business data using a single language and terminology. In [29], we considered the way of transforming SBVR questions into SPARQL queries that can be executed against OWL 2 ontologies obtained from SBVR vocabularies and business rules.

SBVR questions are based on logical projections and are much more comfortable for business people than various query languages that are platform-specific and intended for IT specialists. We can apply several slightly different methodologies for querying: to derive relations using SWRL rules and OWL 2 reasoning tools, and querying using asserted as well as inferred individuals and properties, or formulating direct SPARQL queries. Also, we can store ontology instances in a relational database and apply gradual querying methodology by formulating part of a query in SPARQL and retrieve instances from a relational database using SQL [30].

Here, we will analyze how to translate SBVR questions and business rules into corresponding SWRL and SPARQL expressions. We will concentrate on two aspects: querying facts based on hierarchical and n-ary relations. Taxonomy of semantic relations analyzed by Chaffin [4] and Storey [28] includes seven main relation types: inclusion (class inclusion, meronymic, spatial), possession, attachment, attribution, antonyms, synonyms, and cases. Some meronymic, possession, and attachment relations can arise between objects of the same type. Typical reflexive relations are kinship relations that can comprise complex trees and forests; their analysis requires recursive queries.

N-ary relations have no problems for representing them in UML or SBVR, but they cannot be directly represented in ontology. There are solutions proposed by the W3C for representing n-ary relations in OWL [31] but they are based on the creation of new object types that often seem unnatural for formulating related queries close to natural language. We agree with Hoekstra who argues that the disallowance of n-ary relations is rather an advantage than the drawback of the OWL 2 because n-ary relations can be expressed by binary ones [8]; the presence of n-ary relations demonstrates that the meaning of the corresponding associations is not well-understood. Furthermore, if n-ary relations are un-ordered the semantics of relation can be lost because we may miss the subject of the sentence.

In SBVR and UML, fact type roles and association ends are ordered. However, for keeping the required order of fact type roles in SBVR, one should respectively formulate fact types; in UML, the order of n-ary association ends is set by the order of introducing these ends into a model – this is not always understood be modelers. Our proposed way of representing n-ary relations by binary ones in OWL 2 would be possible if we could define properties for relations (properties of OWL 2 object properties). Then we could distinguish fact type roles of subject and object as roles of a binary relation, and the remaining fact type roles as properties of that relation. In Section IV, we show how to define the SBVR fact types with synonymous forms that allow formulating SSE queries, acceptable for humans, and how to represent these queries in SPARQL by using OWL 2

 ISBN: 978-1-61208-160-1

punning for defining classes and object properties having same IRIs.

III. ANALYZING HIERARCHICAL RELATIONS BETWEEN INDIVIDUAL CONCEPTS

Reflexive relations arise between objects of the same type playing different roles in the relation. For example, in SBVR terminology, fact type "father *has* son" represents associative fact type where fact type roles are played by different persons that are instances of object type "person" (here and in the following, we will use SBVR style "term" for terms, representing noun concepts; *verbs*, representing fact type symbols; "Names" for individuals, and "keywords" for keywords, articles etc making SSE sentences more understandable). We take a genealogy tree as an example of reflexive relations. OWL 2 ontology of a genealogy tree is presented in Figure 1 in UML notation where the association ends correspond to OWL 2 object properties and SBVR fact types e.g., has_son (not fact type roles e.g., "son" as it should be in actual UML models).

In the tree ontology, the asserted classes are "Person", "Marriage" and "Sex"; they have asserted data properties and object properties "has_parent", "has_sex", and "has_marriage". Remaining classes and object_properties can be inferred from OWL 2 axioms and SWRL rules, e.g., "Man" can be defined as "Person" who has sex of male:

```
EquivalentClasses( Man Person
(ObjectHasValue(has_sex value male_sex)))
```

A couple is defined as an object property and can be derived by SWRL rule:

```
has_child(?x,?y),has_child(?z,?y),
  DifferentFrom(?x,?z)→couple(?x,?z)
```

We also use OWL 2 punning and create a class "couple" that is understood as a pair of male and female that have at least one child and, optionally, may be married. Figure 2 presents individuals that are analyzed in the following.

For defining an object property "has_kin" we can use the classical SWRL rules:

```
has_parent(?x,?y)→has_antecedent(?x,?y)
has_parent(?x,?y),has_antecedent(?y,?z)→
has_antecedent(?x,?z)
has_parent(?x,?y)→has_descendant(?y,?x)
has_parent(?x,?y),has_descendant(?y,?z)→
has_descendant(?x,?z)
has_antecedent(?x,?y)→has_kin(?x,?y)
has_descendant(?x,?y)→has_kin(?x,?y)
has_antecedent(?x,?y),has_descendant(?y,?z),
DifferentFrom(?x,?z)→has_kin(?x,?z)
```

That means kin of a person are her antecedents and descendants, as well as descendants of her antecedents except the person herself. We can define rules for all object properties of the genealogy tree in a similar way.

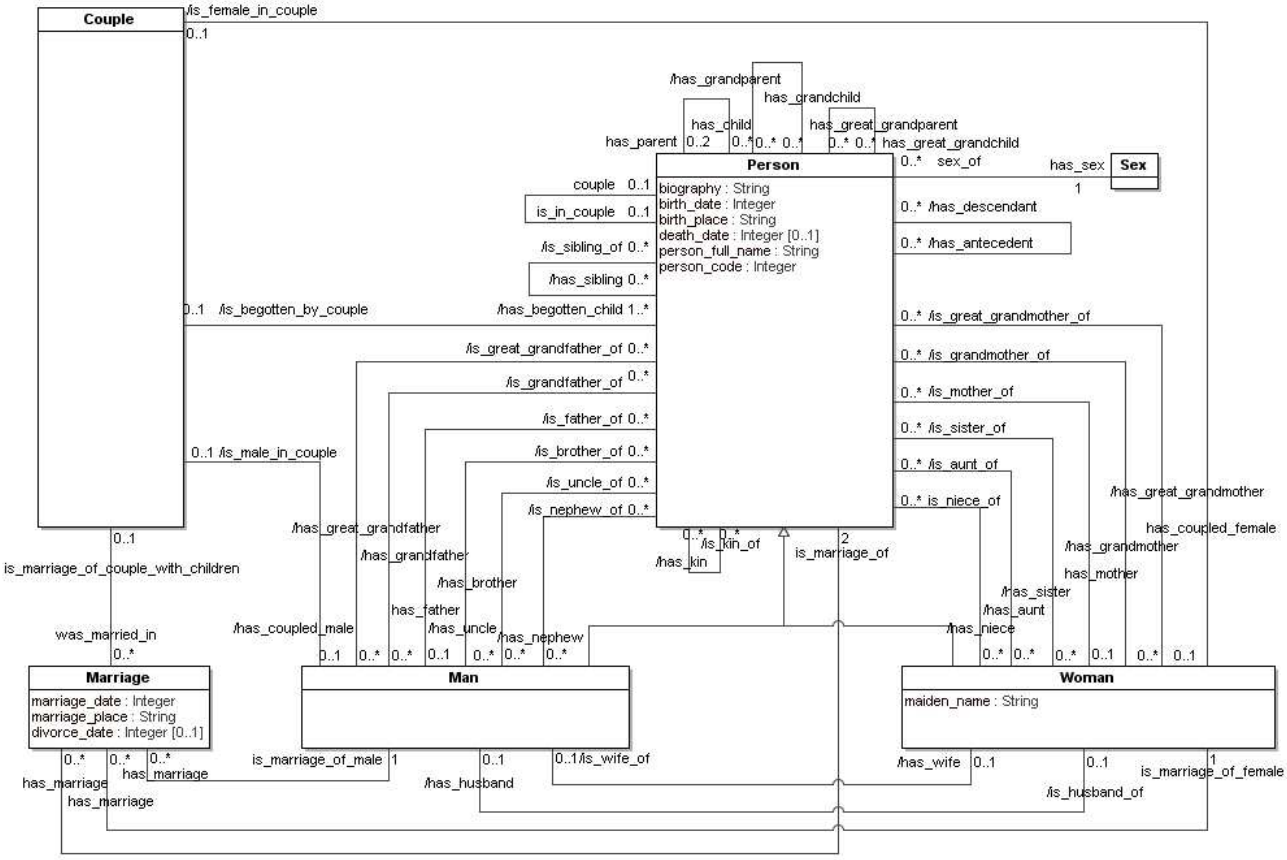

Figure 1. OWL 2 ontology of the genealogy tree in UML notation

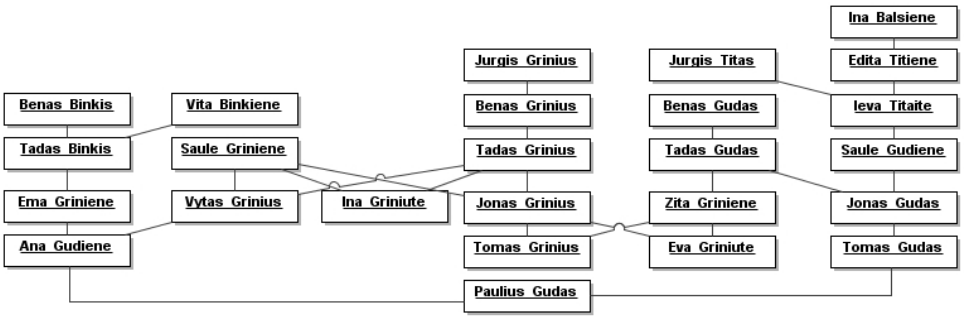

Figure 2. Individuals with "has_parent" links of a genealogy tree ontology in UML notation

Representing SWRL rules in SBVR is straightforward with the use of implication formulation. On the contrary to SWRL, SBVR rules can use disjunctive logical formulations that should be transformed into separate SWRL rules:

It is necessary that person1 has_antecedent person2 if person1 has_parent person2.

…

It is necessary that person1 *has_kin* person2 if
person1 *has_antecedent* person2 or
person1 *has_descendant* person2 or
person1 *has_antecedent* person3 and
person3 *has_descendant* person2 and not
person1 *equals* person2.

Using OWL 2 reasoner (Pellet or Hermit), we can infer kin of all persons and formulate simple SBVR questions as:

"What are *kin_of* person Vita_Binkiene?"

"Are *kin* person Ema_Griniene and Vytas_Grinius?"

that are translated into simple SPARQL queries:

```
SELECT DISTINCT ?kin
{gen:Vita_Binkiene gen:has_kin ?kin},
```

which gives the results presented in Figure 3. For executing queries, we have used the ARQ 2.8.4 query engine that supports SPARQL 1.1 extensions, and Pellet 2.2.2 OWL Reasoner.

```
kin
<http://Data.Genealogy_tree/Tadas_Binkis>
<http://Data.Genealogy_tree/Ema_Griniene>
<http://Data.Genealogy_tree/Ana_Gudiene>
<http://Data.Genealogy_tree/Paulius_Gudas>.
```

Figure 3. Results of the query "What are *kin_of* person Vita_Binkiene?"

The second query is of ASK type (the "kin" is the synonymous noun form of "has_kin"):

```
ASK{gen:Ema_Griniene
gen:has_kin gen:Vytas_Grinius}
```

and gives the result `false`.

Without the use of reasoner, formulating SPARQL queries from SBVR questions can become complicated as SPARQL recursive queries are problematic to identify from SBVR rules:

```
select distinct ?kin{{select ?antecedent{
gen:Vita_Binkiene gen:has_parent* ? antecedent .
not exists {? antecedent gen:has_parent ?x}}}
?kin gen:has_parent* ? antecedent .
filter(?kin != gen:Vita_Binkiene)}
```

Another way of transforming SBVR questions into SPARQL is to formulate derivation rules in CONSTRUCT query form [23]. However, CONSTRUCT queries inefficiently work with large rule sets. Therefore we transform SBVR rules into separate CONSTRUCT queries that give results in separate triple graphs, and apply Jena Union function that dynamically relates result graphs through their common elements. Finally, we execute SELECT query in the united graph:

```
CONSTRUCT {?x gen:has_antecedent ?y .}
WHERE{?x gen:has_parent ?y .}
...
CONSTRUCT{?x gen:has_kin ?y .}
WHERE{?x gen:has_antecedent ?y .}
CONSTRUCT{?x gen:has_kin ?y .}
WHERE{?x gen:has_descendant ?y .}
CONSTRUCT{?x gen:has_kin ?z .}
WHERE{?x gen:has_antecedent ?y .?y
gen:has_descendant ?z .filter(?x != ?z)}
SELECT DISTINCT ?kin{gen:Vita_Binkiene
gen:has_kin ?kin}
```

The SELECT gives the same results as in Figure 3.

IV. DEFINING N-ARY RELATIONS IN OWL 2

SBVR metamodel allows formulating fact types with more than two fact type roles and preserves the ordering of these roles. However, for further retaining the meaning of relations it is desirable to express them as binary relations because in the fact type there always are two main fact type roles – subject and object; the remaining fact type roles mean either properties of the fact type relating subject and object, either properties of object. We should formulate fact types in such a way that it would be possible to identify what fact type roles mean properties of fact type, and what ones mean properties of object. That is, instead of defining SBVR fact type as

"person *works_in* organization in position from date till_date",

it is desirable to represent it in such a way:

"person *works_in* organization from_date till_date",

Synonymous form: position,

"person *occupies_position_of* position",

"position *performs_office_of* office".

Here the last fact type is introduced for representing object type "office" that is a role type of the role "position". Figure 4 presents example ontology for modeling properties of relation in OWL 2. Role "Position" is explicitly presented as a class. Dependency with stereotype <<EquivalentClasses>> means that EquivalentClasses axiom

is defined for classes "works_in" and "Position". We cannot define properties of relations (i.e., object properties) in OWL 2, but we can use punning and define a class with the same name as the object property "works_in". For flexibility of formulating propositions and questions, we create the additional class "Position" equivalent to class "works_in", and attribute properties of the relation for that class.

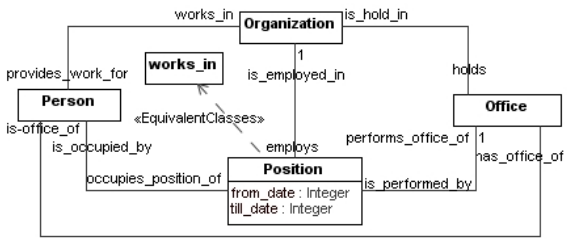

Figure 4. Example of modeling properties of relations in UML

Guizardi and Wagner suggested the similar solution [7] (without the role type) for representing properties of relations in UML despite that UML has the association class for this purpose. They have criticized the ambiguity of UML association class and proposed to create an association for representing a formal relation, and a separate class for materializing that relation.

Having such ontology, we can simply ask

"What are organizations that person Jonas_Grinius *works_in* from_date 2000-01-01 till_date 2003-01-01?"

Here we cannot use current reasoners (Pellet, Hermit) because they do not derive relation properties on the base of punning but we can attain these properties using SPARQL. For this, we should supplement the question with a business rule

"It is necessary that person *works_in* organization from_date till_date if the person *occupies_position_of* position from_date till_date and the position *is_employed in* the organization."

Then we instantiate the rule for person "Jonas_Grinius" and transform the question and the rule into the SPARQL query (!bound is added to optional variables):

```
select distinct ?org{
org:Jonas_Grinius org:works_in ?org.
org:Jonas_Grinius
org:occupies_position_of ?position .
?position org:is_employed_in ?org .
?position org:from_date ?from .
optional {?position org:till_date ?till} .
filter (?from <= "2003-01-01"^^xsd:date ) .
filter (?till >= "2000-01-01"^^xsd:date ||
!bound(?till))}
```

For the given ontology example of organizations and persons we also can define the OWL 2 object property chain:

```
SubObjectPropertyOf(ObjectPropertyChain
(occupies_position_of performs_office_of)
has_office_of)
```

that is represented as SBVR structural business rule:

It is necessary that person *has_office_of* office if the person *occupies_position_of* position and the position *performs_office_of* the office.

Current OWL 2 reasoners understand object property chains. As previously, we will consider two cases: 1) when we use a reasoner and formulate a simple question; 2) when we do not use a reasoner and formulate a question together with derivation rule.

Using reasoner, we can give the SBVR question

"What person *has_office_of* Professor?"

that transforms into the SPARQL query:

```
SELECT ?persons
{?persons org:has_office_of gen:Professor}
```

SBVR question without reasoning

"What person *occupies_position*

that *performs_office_of* Professor?"

transforms into SPARQL query:

```
SELECT ?persons{?persons
org:occupies_position_of ?position .
?position org:performs_office_of
org:Professor}
```

V. CONCLUSIONS ANF FURTHER WORKS

The paper presented some formulations of SBVR fact types and questions for recursive and n-ary relations, together with representing them in OWL 2, SWRL and SPARQL, and executing obtained queries in current ontology reasoning tools and SPARQL engines. For analyzing recursive relations, we represent them by SBVR derivation rules and propose two scenarios for transforming them: 1) using SWRL rules, when we obtain simple SPARQL queries but should apply OWL 2 reasoner in addition to SPARQL engine; 2) using sequences of SPARQL CONSTRUCT queries. Both cases are practicable in real life applications. For analyzing n-ary relations, we propose to represent them by using SBVR synonymous forms. Transformation of these forms into OWL 2 results in introducing a new class and a binary relation with the same name (allowed by OWL 2 punning) for defining the main relation between subject and object; and object properties of that class representing remaining roles of n-ary relation. Such representation allows avoiding unnatural names that appear in traditional solution.

The limitations of the approach are in the fact that OWL 2, even supplemented with SWRL, is not enough to model the complete semantics of SBVR vocabularies, and, consequently, SPARQL is not enough to model SBVR questions. From the other side, it still is not clear how to represent some OWL or SPARQL constructs in SBVR in a natural way. Our future work will focus on analysis of more patterns for improved flexibility of SBVR questions translatable to SPARQL queries as well as providing experiments for integrating the proposed solutions into the realistic enterprise context. Also, we have some initial prototype of SBVR editor for Lithuanian language and are willing to work for embodying the SBVR Structured Lithuanian.

REFERENCES

[1] L. Bodenstaff, P. Ceravolo, R. E. Damiani, C. Fugazza, K. Reed, and A. Wombacher, "Representing and Validating Digital Business Processes", in Web Information Systems and Technologies, LNBIP, vol. 8(1), 2008, pp. 19–32.

 ISBN: 978-1-61208-160-1

[2] J. Cabot, R. Pau, and R. Raventos, “From UML/OCL to SBVR specifications: A challenging transformation”. Information Systems, 2008, pp. 1–24.

[3] L. Ceponiene, L. Nemuraite, and G. Vedrickas, “Semantic business rules in service oriented development of information systems”, in Information Technologies' 2009: proceedings of the 15th International Conference on Information and Software Technologies, IT 2009, Kaunas, Lithuania, April 23-24, 2009, pp. 404–416.

[4] R. Chaffin and D. J. Herrman, “The Nature of Semantic Relations: A Comparison of Two Approaches”, in Relational Models of the Lexicon: Representing Knowledge in Semantic Networks. M. Evens, Ed. New York: Cambridge University Press, 1988, pp. 289–334.

[5] B. Demuth and H. B. Liebau, “An Approach for Bridging the Gap Between Business Rules and the Semantic Web”, in Advances in Rule Interchange and Applications, LNCS, vol. 4824, 2009, pp. 119–133.

[6] S. Goedertier and J. A. Vanthienen, “A Vocabulary and Execution Model for Declarative Service Orchestration”, in Business Process Management Workshops, LNCS, vol. 4928, 2008, pp. 496–501.

[7] G. Guizzardi and G. Wagner. „What’s in a Relationship: An Ontological Analysis“, in Conceptual Modeling – ER 2008. LNCS, vol. 5231, 2008, pp. 83–97.

[8] R. Hoekstra. Ontology Representation Design Patterns and Ontologies that Make Sense. SIKS Dissertation Series No. 2009-15, SIKS, Dutch, 2009.

[9] A. Kamada and M. Mendes, “Business Rules in a Service Development and Execution Environment”, in Eleventh International IEEE EDOC Conference Workshop (EDOCW'07), 2007, pp. 1366–1371.

[10] E. Kaufmann and A. Bernstein, “Evaluating the Usability of Natural Language Query Languages and Interfaces to Semantic Web Knowledge Bases”. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 8, 2010, pp. 1–23.

[11] M. Kleiner, P. Albert, and J. Bézivin, “Parsing SBVR-Based Controlled Languages”, in Model Driven Engineering Languages and Systems, LNCS, vol. 5795, 2009, pp. 122–136.

[12] M. Kriechhammer, “Querying Systems for business models” 2006. Available from: http://www.kriechhammer.com /?English_Portfolio:my_Documents:Finals.[Accessed 06 May 2011].

[13] I. Lemmens, M. Nijssen, and S. Nijssen, “A NIAM2007 Conceptual Analysis of the ISO and OMG MOF Four Layer Metadata Architectures”, in On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops, LNCS, vol. 4805, Germany: Springer Berlin/Heidelberg, 2007, pp. 613–623.

[14] M. H. Linehan, “Ontologies and Rules in Business Models”, in Proc. 11th Int. EDOC Conference Workshop (EDOC '07), 2007, pp. 149–156.

[15] M. H. Linehan, “Semantics in Model-Driven Business Design”, in Proc. 2nd Int. Semantic Web Policy Workshop (SWPW'06), 2006.

[16] A. Marinos and P. Krause, “An SBVR Framework for RESTful Web Applications”, in Rule Interchange and Applications, LNCS, vol. 5858, Germany: Springer Berlin/Heidelberg, 2009, pp. 144–158.

[17] L, Nemuraite, T. Skersys, A. Sukys, E. Sinkevičius, and L. Ablonskis, “VETIS tool for editing and transforming SBVR business vocabularies and business rules into UML&OCL models”, in Information Technologies' 2010: proceedings of the 16th International Conference on Information and Software Technologies, IT 2010, Kaunas, Lithuania, April 21-23, 2010, pp. 377–384.

[18] J. Nenortaite and R. Butleris, “Improving Business Rules Management through the Aplication of Adaptive Business Intelligence Technique”. Information Technology and Control, vol. 38(1), 2009, pp. 21–28.

[19] S. Nijssen, **“**SBVR: Semantics for Business”. Business Rules Journal, vol. 8(10), Oct. 2007, available from: http://www.BRCommunity.com/a2007/b367.html [Accessed 06 May 2011].

[20] S. Nijssen, “SBVR ~ Ground Facts and Fact Types in First-Order Logic”, Business Rules Journal, vol. 9(1), Jan. 2008, Available from: http://www.BRCommunity.com/a2008/b386.html [Accessed 06 May 2011].

[21] OMG, 2008. Semantics of Business Vocabulary and Business Rules (SBVR). Version 1.0. December, 2008, OMG Document Number: formal/2008-01-02.

[22] B. Paradauskas and A. Laurikaitis, “Business Knowlegde extraction using program understanding and data analysis techniques”, in Information Technologies' 2009: proceedings of the 15th International Conference on Information and Software Technologies, IT 2009, Kaunas, Lithuania, April 23-24, 2009, 337–354.

[23] A. Polleres, F. Scharffe, and R. Schindlauer. “SPARQL++ for Mapping between RDF Vocabularies”, in OTM'07 Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS – Volume Part I, Springer-Verlag Berlin, Heidelberg, 2007.

[24] E. Prud‘hommeaux and A. Seaborne, A. SPARQL Query Language for RDF. W3C Recommendation. 15 January 2008. Available from: http://www.w3.org/TR/rdf-sparql-query/ [Accessed 06 May 2011].

[25] A. Razavi, A. Marinos, S. Moschoyiannis, and P. Krause, “RESTful Transactions Supported by the Isolation Theorems”, in Web Engineering, LNCS, vol. 5858, Germany: Springer Berlin/Heidelberg, 2009, pp. 394–409.

[26] S. Spreeuwenberg and H. K. Anderson, “SBVR's approach to controlled natural languages”, in Workshop on Controlled Natural Language (CNL 2009), 2009.

[27] S. Spreeuwenberg and R. Gerrits, “Business Rules in the Semantic Web, are there any or are they different?”, in Reasoning Web. Springer Berlin/Heidelberg, Germany, LNCS, vol. 4126, 2006, pp. 152–163.

[28] C. Storey, “Understanding Semantic Relationships”. VLDB Journal, vol. 2, 1993, pp. 455–488.

[29] A. Sukys, L. Nemuraite, B. Paradauskas, and E. Sinkevicius “Querying ontologies on the base of semantics of business vocabularies and business rules”, in Information Technologies' 2011: proceedings of the 17th international conference on Information and Software Technologies, IT 2011, Kaunas, Lithuania, April 27-29, 2011. Kaunas, 2011, pp. 247–254.

[30] E. Vysniauskas, L. Nemuraite, and A. Sukys, “A hybrid approach for relating OWL 2 ontologies and relational databases”, in Perspectives in Business Informatics Research: proceedings of the 9th international conference, BIR 2010, Rostock, Germany, September 29 – October 1, 2010, Berlin-Heidelberg-New York, Springer, 2010, pp. 86–101.

[31] W3C. Defining N-ary Relations on the Semantic Web. W3C Working Group Note, Apr. 2006. Available from: http://www.w3.org/ TR/swbp-n-aryRelations [Accessed 06 May 2011].

[32] W3C. SPARQL 1.1 Query Language. W3C Working Draft 14 October 2010. Available from: http://www.w3.org/TR/sparql11-query/ [Accessed 06 May 2011].

 ISBN: 978-1-61208-160-1

Enterprise Knowledge Modeling and Data Mining Integration

Auksė Stravinskienė
Kaunas Faculty of Humanities of
Vilnius University
Kaunas, Lithuania
Aukse.Stravinskiene@khf.vu.lt

Saulius Gudas
Kaunas Faculty of Humanities of
Vilnius University
Kaunas, Lithuania
Saulius.Gudas@khf.vu.lt

***Abstract*—The control view-based approach to integration of enterprise modeling and data mining is the focus of the paper. The analysis and evaluation of enterprise modeling methods and languages (ARIS, UEML, BPMN and Value Chain Model) for elicitation of enterprise knowledge components are presented. The knowledge components of control view-based Enterprise Meta-model are discussed. The principles of identification of enterprise knowledge components and their integration to data mining process are defined.**

Keywords-Enterprise Modeling; Data mining; Control view; Enterprise Meta-model; Knowledge

I. INTRODUCTION

The rapid development of Business Intelligence (BI) systems requires elicitation of enterprise management experience as knowledge and using it for development of Intelligent Information Systems [1]. It is important to develop an adequate business model that matches the organization's management information interactions.

The enterprise modeling for information systems (IS) engineering is aimed at to describing the scope of the organization's management activities – knowledge and data transformations in enterprise management systems. Enterprise model is required to improve the company's performance, describing the operational sequences, responsibilities and relationships as well as to identify must systems that duplicate each other or are otherwise inappropriate [2].

Even though enterprise models are different, they all serve the purpose of developing enterprise models and methods, which could be used for the IS development as knowledge-based integration of data mining, online analytical processing (OLAP). Using such models it is easier to understand enterprise process, enterprise knowledge and structure, and to improve the enterprise efficiency.

Enterprise models for information systems engineering are developed using DFD, ARIS, UML, IDEF3, IDEF0 notations [2,8,9,10]; especially, BPMN and UEML notations are effective for enterprise modeling.

It is important to identify the enterprise management knowledge which can be useful to improve enterprise efficiency, to help user and to manage data mining process.

There are two types of models for IS engineering: a) enterprise models that describe empirical information about domain and b) enterprise management models that verify empirical information against formal theories or predefined structures [3]. Empirical enterprise models as well as enterprise management models can be developed using all previously mentioned business process modeling notations. A specific characteristic of the Enterprise management models is that they are based on obligatory information relationships of enterprise components. For instance, they form a feedback loop between the components of managed process and enterprise management function.

Information need of enterprise activities depend on the certain patterns of enterprise management process, identified in the Enterprise management model [4]. The knowledge-based approach to IS development is based on the Enterprise Meta-model [4] and co-relates with knowledge modeling and discovered of enterprise knowledge components. **T**ypes of enterprise model components and their interactions are defined in Enterprise Meta-model [4]. Therefore, an Enterprise Meta-model component could be used for acquisition of enterprise knowledge components.

These principles of enterprise management modeling could be used for improvement data mining process. Reasonable selection of a data mining method and an algorithm is important to successful implementation of the process of analysis and the final result of data mining.

Enterprise and enterprise management models are composed of components, which will be regarded as sources of enterprise knowledge.

On the basis of knowledge-based engineering principles it is essential to identify the enterprise's knowledge components that are involved in enterprise management.

This paper analyses EMM, ARIS, UEML, BPMN enterprise modeling methods and identifies knowledge-based components [2,8,9,10]. A control view-based approach to identification of knowledge components and their integration to data mining process are presented in the paper.

 ISBN: 978-1-61208-160-1

The paper structured into 3 parts. First, Enterprise knowledge and data mining integration are defined. Second, Knowledge-based analysis of Enterprise modeling methodology is described. Third, Enterprise knowledge model's integration into data mining is analyzed.

II. ENTERPRISE KNOWLEDGE AND DATA MINING

Data mining is an iterative process of searching in large amounts of data or in large data repositories for hidden, previously unknown and potentially useful information, knowledge, a set of patterns and relationships between data. Data mining is a process (based on the CRISP-DM model) composed of consistent set of six stages [5]: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, Deployment.

These six data mining stages construct a single cycle of process. In order to gain better data mining results, data mining process could be aligned with Enterprise management modeling [6]. We propose to integrate the knowledge base in a data mining process, on the basis of which the data mining process could be controlled.

The component of Enterprise management model could be used in data mining such as Enterprise knowledge held by an expert. Its use can develop a more rational data mining queries. We propose to use the Knowledge base in which Enterprise management knowledge would be collected. More accurate and structurally interesting information can be extracted from Knowledge base by adding business rules, knowledge of experience about management functions.

Knowledge-based management model (Figure 1) is being established according to this principle: Enterprise management is structured as interactions of Management functions and Enterprise processes [6]. This Knowledge-based management model illustrates the interaction between Enterprise processes (Pj) and Management functions (Fi), which serves for the Knowledge–based Enterprise's information system. The model (Fi x Pj) consists of Enterprise's information system, Management functions (F1xFi), Business goals (G), Enterprise processes (P1xPj).

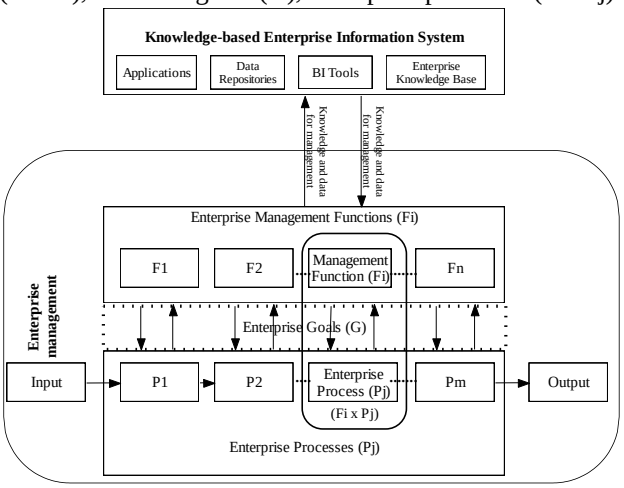

Figure 1. Knowledge – based management as interactions of Management functions and Enterprise processes

Knowledge-based Information System consists of Data Repositories, Applications (Management IS), as well as of Enterprise Knowledge base, which supports Applications and BI Tools with knowledge components.

The influence of Enterprise Goals to Enterprise management function is critically important as the final result of management depends on the Enterprise Goals.

Control view-based Enterprise management structure is presented in Figure 2.

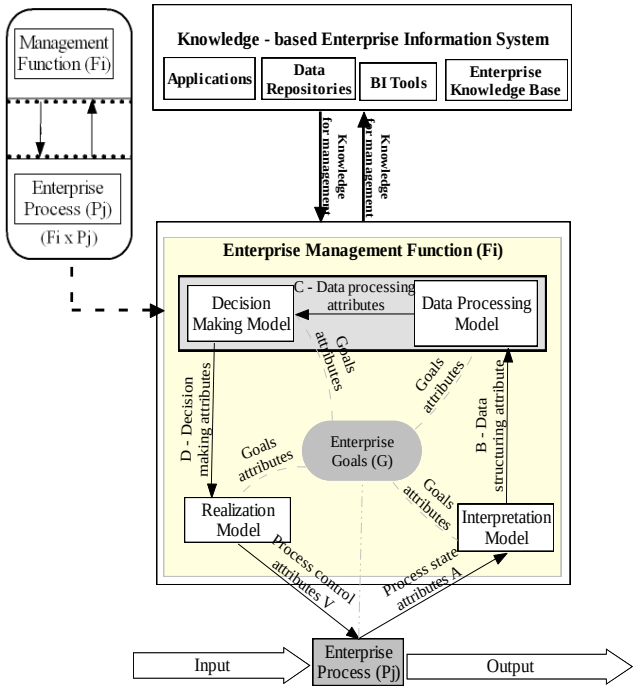

Figure 2. Control view-based structure of Enterprise management function Fi (Gudas S., Skersys T., Lopata A., 2004)

Enterprise management is a model identifying Enterprise components, components of Enterprise management (Data, Knowledge, Goals) and their interactions [6]. "EMC (Elementary Management Cycles) is formalized description of the interaction of Process and Function – as two core components of Enterprise from the control point of view"[7].

Application of this model is important because this model describes the essential elements of Enterprise management and control function and their interactions. Besides, information flows are identified (Figure 2). The model reflects the interaction of management function and Enterprise process in which the elementary management activities (Interpretation, Data processing, Decision making, Realization of decision) compose a cycle– an Elementary Management Cycle (EMC).

Decomposition of Knowledge management system interactions with Business Intelligence Tools and Applications are presented in Figure 3.

Figure 3 displays a detailed scheme of Enterprise management function integration into data mining (aligned with Figure 1 and 2), in which it is clearly shown what interfaces there are between Business Intelligence Tools,

 ISBN: 978-1-61208-160-1

Applications, Data repositories and Enterprise knowledge base and which data is integrated from the Enterprise Knowledge base.

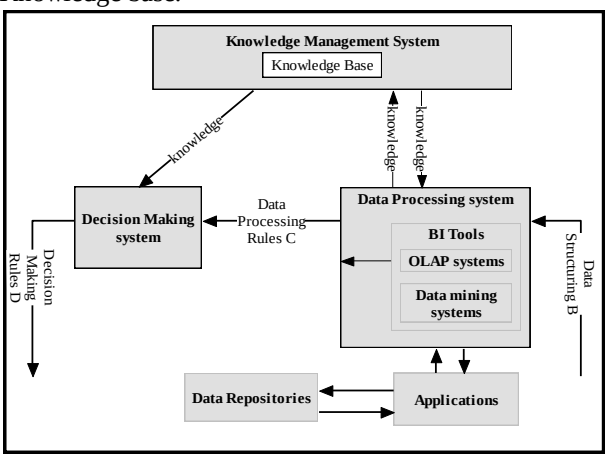

Figure 3. Integration of Knowledge management system and data mining

Knowledge, which is stored in Knowledge Base, is going to be used in the process of data mining for algorithm modification.

III. KNOWLEDGE-BASED ANALYSIS OF ENTERPRISE MODELING METHODOLOGY

This section provides the analysis of UEML, ARIS, BPMN and EMM enterprise modeling methodologies and languages from the aspect of knowledge. Components of structure of an enterprise model are described and knowledge components of a certain method are identified.

A. *UEML Enterprise modeling language*

Unified Enterprise Modeling Language (UEML) is a language of Enterprise processes' modeling, which is aimed at facilitating the integration of different modeling languages in a company [8]. UEML model stores Enterprise knowledge that is shown in Figure 4.

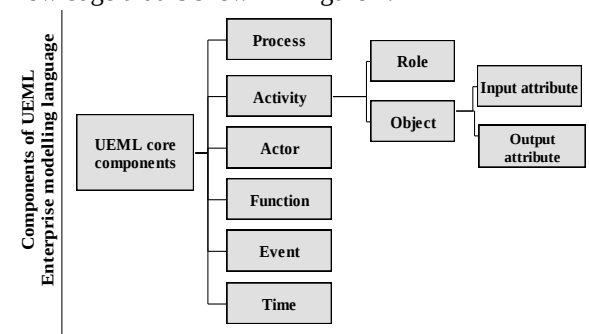

Figure 4. The major components of the UEML

Figure 4 shows the composition of components of the UEML model: process, activity, actor, function, event, time, role, object, input and output. Function is defined as a component, which consists of processes and interactions of activities (including feedback loop). At the Knowledge base, each component stores a certain attribute or value that identifies it. The structure of UEML model can be described by these constructs:

UEML= {Process, Activity, Actor, Function, Event, Role, Enterprise Object (Product, Order, Resourse)}

B. *ARIS Enterprise modeling methodology*

Architecture of Integrated Information Systems (ARIS) - method of Enterprise modeling [9]. The main concepts of ARIS are a) the architecture for describing business processes; b) modeling tools for Event-Driven Process; c) comprehensive computer-aided Business process management. This methodology includes design, implementation, optimization and controlling of Business process. The uniqueness of this methodology is modeling from the control view point. Figure 6 shows a hierarchical structure of components of the ARIS method, in which constituent parts of the method are reflected.

ARIS= {Process, Activity, Objectives, Function, Control, Information}

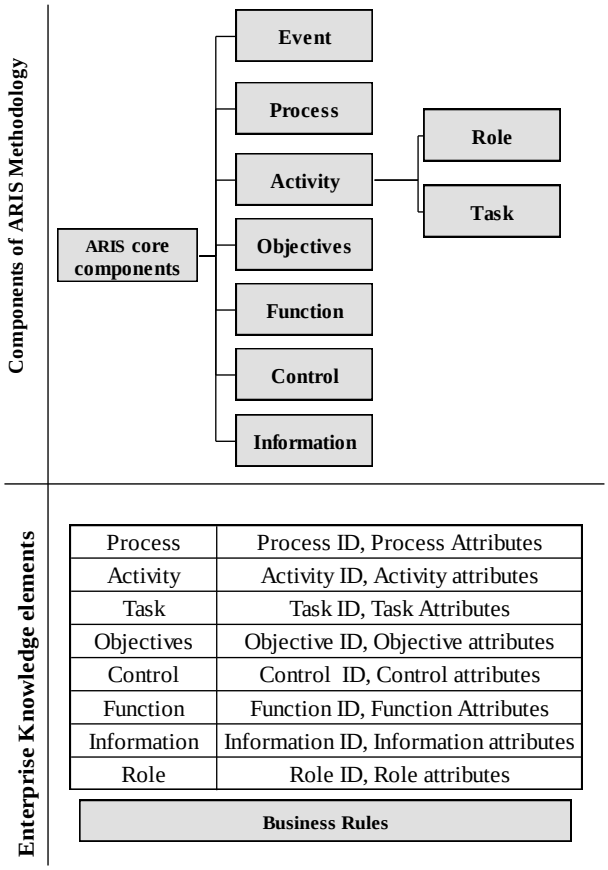

Figure 5. The major components of the ARIS method

C. BPMN Enterprise modeling notation

Business Process Management Notation (BPMN) is a notation for the formation of business processes' flow models (Figure 6). It is coordinated by Object Management Group (OMG) standards' organization that focuses on improving business processes. The basis of this methodology is the possibility for representatives of different fields of business to understand business processes that take place in an organization and to reveal them in a single modeling notation [11]. Composition of BPMN model can be described by constructs that are presented in Figure 6:

BPMN= {Pool, Process, Task, Activity, Gateway, Lane, Association, Event, Data Object}

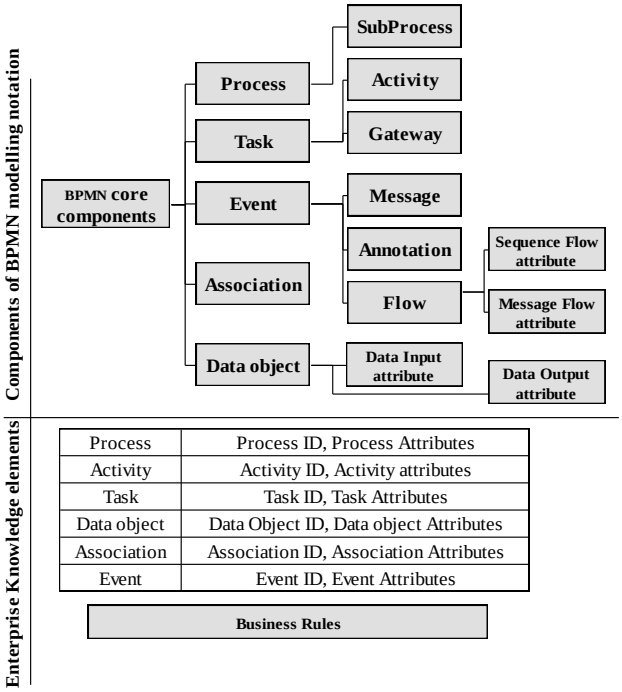

Figure 6. The major components of the BPMN model

D. Enterprise Meta-model

EMM is a model of managed processes based on the structure of Elementary management cycle (EMC) that can be used for Knowledge-based systems modeling [4]. The main principle of this model's structure is the evaluation of an aspect of Enterprise's management by modeling informational interactions of Enterprise processes and Enterprise management functions.

The control view-based approach to Enterprise modeling results in the Enterprise Meta-model (EMM)[4]. The EMM model is a formal description of the Enterprise management knowledge, in which major types of Enterprise components, their types of interactions are reflected (Figure 7).

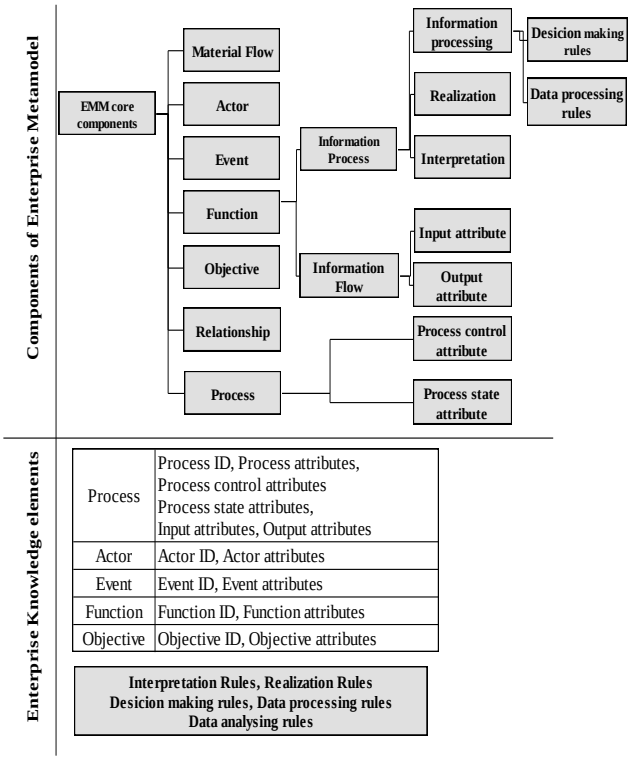

Figure 7. Components of Enterprise Meta-model (EMM) (Gudas S. et al., 2005)

E. Comparison of Enterprise Modeling Methods

Table 1 show a summary of the Enterprise modeling methods (Enterprise Meta-models) and it's, which demonstrates what knowledge elements are identified by each Enterprise modeling method. Each component expresses an aspect of Enterprise knowledge, which can be accumulated in the Knowledge base and developed on the basis of such an EMM. It indicates which Enterprise model accumulates more knowledge of the organization. A comparative analysis shows that Enterprise Meta-model (EMM) accumulates more knowledge components about the activities of the organization (Figure 7) as it models Enterprise management function in detail because of the control-view approach [4].

 ISBN: 978-1-61208-160-1

TABLE I. COMPARISON OF ENTERPRISE MODELING METHODS

Components of Enterprise Knowledge		ARIS	UEML	BPMN	EMM
Process	Process ID	+	+	+	+
	Process State Attributes				+
	Process Control Attributes				+
Function	Function ID	+			+
	Process State Attributes A				+
	Process Control Attributes V				+
	DP Input Attributes B				+
	DP Output AttributesC				+
	DM Output AttributesD				+
	IN Rules				+
	DP- Data Processing Rules				+
	DM -Decision Making Rules				+
	Decision Realization Rules				+
	Data Analysis Rules				+
Event		+	+	+	+
Role		+	+		+
Actor			+		+
Activity		+	+	+	+
Goal, Objective		+		+	+
Control		+			+
Time			+	+	
Material Flow				+	+
Input		+	+	+	+
Output		+	+	+	+
Object			+		
Interpretation Rules					+
Realization					+
Decision control					+
Data structuring					+
Business Rules		+	+	+	

Table 1 lists the components of different Enterprise modeling approaches, indicates Enterprise knowledge components and attributes. After the analysis of the models and their constituent components we may presume that the EMM model retains more Enterprise knowledge components that are needed in order to enhance the IT based organization's management functions.

IV. ENTERPRISE KNOWLEDGE MODEL'S INTEGRATION INTO DATA MINING

Enterprise knowledge model is a model of information management processes, identifying Enterprise components and their interactions, components of Enterprise management (data, knowledge, goals) and their interactions [6]. Application of this model for the development of a Knowledge-based information system is important because it reflects the essential elements of the model and their interactions and displays information flows (Figure 5). The element of knowledge management is important in this model. Adding certain rules, knowledge of experience to functions performed by process, more accurate and structurally interesting information can be extracted from data sets.

Figure 8 represents a diagram of entities' connections of Enterprise management model, which consists of the following components: Process, Function, Goals, Interpretation rules, Realization rules, Decision making rules, Data processing rules and Interactions between them. This is a transformed diagram of Enterprise Meta-models defined in Figures 2 and 3, the aim of which is to show the interfaces of Enterprise knowledge with process and function.

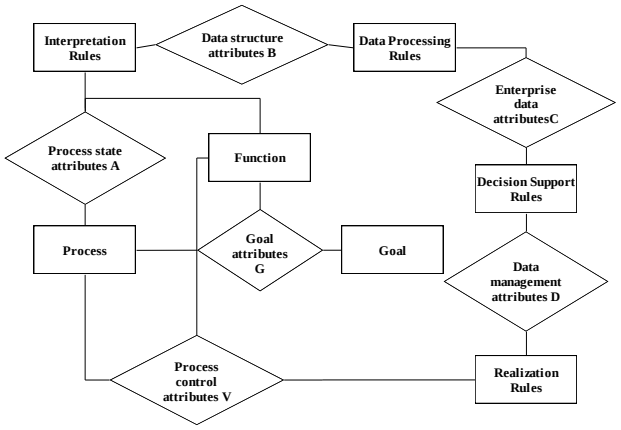

Figure 8. Enterprise Knowledge-based model - Entity relationships diagram (Gudas S., 2004)

In particular, Enterprise model helps to define and optimize processes in a company. Sequence of events, their interactivity, responsibility and data sources are clearly defined in the model. The composed Enterprise model allows the user to clearly imagine the data mining process and to adapt it to their own company easily and efficiently.

Components of knowledge that can be expressed in a form of Enterprise rules are accumulated in Knowledge base. In this base, rules are stored independently from software code, thus, ensuring the flexibility of a system and the opportunity to quickly adapt the functionality of the system under the changed conditions, but without changing the software code. Template of Enterprise rules can be expressed as follows:

IF <Variable><Condition><Value> THEN
<Variable><Condition><Value>

Figure 9 depicts mapping of the structure of Enterprise Meta-model into data mining process. Data mining process consists of several stages: creation of an OLAP cube, which will be used for data analysis; making a cube scheme; creation of a new data mining model; creation of additional fields, columns for implementation of a data mining algorithm; implementation of data mining process [11]. The table “Enterprise Meta-model” (Figure 9) provides knowledge components of Enterprise Meta-model that can

be transmitted to a certain stage of data mining process and monitors its implementation. Knowledge bases and interaction between database and data mining process are also highlighted.

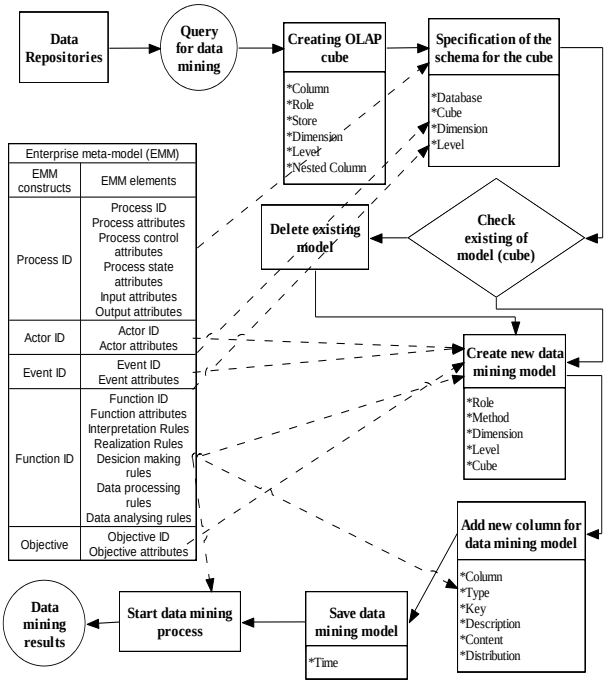

Figure 9. Mapping of EMM components into data mining process steps

A few types of Business rules are identified by EMM (IN rules, DP rules, etc.) usually have a declaratory form, note certain events and express them in conditional sentences. Figure 9 presents the model, in which the components of Enterprise management knowledge can be used for a specific data mining process phase. They may be useful in choosing the data mining algorithm, through which the data will be analyzed. According to the formed rules that would define the rules for algorithm model selection, the stage of model selection would be controlled.

Such Enterprise business rules, accumulated in Knowledge base, can be used in the process of data mining: Data processing, Realization, Decision making, and Interpretation rules.

Interpretation rules may be useful in shaping dimensions and determining the variables needed for data mining model, by determining roles and selecting methods for data analysis. Data processing rules will allow controlling data selection for data mining. Realization rules can be used in choosing the format of displaying results gained during the data mining process.

V. CONCLUSIONS

In this paper the control view-based approach to integration of data mining and Enterprise knowledge modeling was discussed.

The analysis of Enterprise modeling methods and languages (ARIS, UEML, BPMN and detailed Value Chain Model) was performed for elicitation of Enterprise knowledge components.

The control view-based Enterprise Meta-model EMM (Figure 7) is selected as a rational alternative for Enterprise management knowledge modeling. A comparative analysis of these Enterprise modeling methods in terms of knowledge components has shown that EMM is distinguished from other Enterprise modeling methods by a detailed structure of identified knowledge.

Our analysis has shown that it is possible to integrate a Knowledge base, which is designed for EMM, into data mining process to streamline the implementation of a data mining algorithm (Figure 9). Such a data mining system is enhanced by Enterprise knowledge subsystem, and can give additional semantics and functionality for experts to mine more accurate and relevant Enterprise data and knowledge components.

REFERENCES

[1] Josef W. Seifert. Data Mining: An Overview, CRS Report for Congress, Received through the CRS Web. Order Code RL31798, 2004

[2] Saulius Gudas. Žiniomis grindžiamos IS inžinerijos metodų principai. International conference, IT-2005, KTU, Kaunas, 2005

[3] Saulius Gudas, Tomas Skersys and Audrius Lopata. Framework for Knowledge-Based IS Engineering. ADVIS 2004: 512-522

[4] Saulius Gudas, Tomas Skersys and Audrius Lopata. Approach to Enterprise Modeling for Information Systems Engineering. Informatica, Vol. 16, No. 2, Institute of Mathematics and Informatics, Vilnius, 2005, pp. 175-192

[5] The CRISP-DM consortium (2011-06-25) Process Model. The CRISP-DM project. <http://www.crisp-dm.org/Process/index.htm>

[6] Saulius Gudas, Tomas Skersys and Audrius Lopata. (2004). Framework for knowledge-based IS engineering. Advances in Information Systems: 3rd international conference, ADVIS 2004, Izmir, Turkey, October 20-22, 2004: Proceedings. Berlin, Springer, pp. 512-522. ISBN 3-540-23478-0

[7] Saulius Gudas and Rasa Brundzaitė. Decomposition of the Enterprise Knowledge Management Layer, 2006 Seventh International Baltic conference on databases and information systems-proceedings, pp. 41-47, 2006 ISBN 1-4244-0345-6

[8] François Vernadat. UEML: Towards a Unified Enterprise modeling language. International Conference on Industrial Systems Design, Analysis and Management (MOSIM'01), Troyes, France, 2001

[9] IDS Scheer. ARIS Platform Products (2011-06-25) http://www.bptrends.com/publicationfiles/04-08-PR-BPM-Tools Report-IDS Scheer.pdf

[10] OMG Modeling and Metadata Specifications (2011-06-28) http://www.omg.org/technology/documents/modeling_spec_catalog.htm

[11] Daniel T. Larose. Data mining methods and models, IEEE Computer Society Press, pp. 344, ISBN-13 978-0-471-66656-1., 2006

 ISBN: 978-1-61208-160-1

A Business Intelligence Infrastructure Supporting Respiratory Health Analysis

R. Dinis, A. Ribeiro, Maribel Y. Santos
Centro Algoritmi, Universidade do Minho
Guimarães, Portugal
{pg15527, pg15287}@alunos.uminho.pt,
maribel@dsi.uminho.pt

Jorge Cruz*, A. Teles de Araújo**
*Faculdade de Medicina de Lisboa
**Fundação Portuguesa do Pulmão
Lisboa, Portugal
costacruzjorge@gmail.com, artur@telesdearaujo.com

***Abstract*—Business Intelligence Systems are being designed and implemented to support data analysis tasks that help organizations in the achievement of their goals. These tasks are accomplished using several technologies aimed to store and analyze data. This paper presents the particular case of the design and implementation of a business intelligence system to support health care specialists in the analysis and characterization of symptoms related with the chronic obstructive pulmonary disease. For this specific application domain, a data mart model is proposed, implemented and loaded, allowing the analysis of the available data using on-line analytical processing technology and a spatial data mining algorithm. The results obtained so far are promising, demonstrating the usefulness of the proposed business intelligence approach to characterize the key factors in the comprehension of the disease under analysis.**

Keywords-business intelligence; data mining; on-line analytical processing; chronic obstructive pulmonary disease

I. INTRODUCTION

The collection and storage of huge amounts of data is increasing as organizations need to analyze these data and identify useful patterns, trends or models that support the decision making process. Independently of the application domain, this is the reality of nowadays organizations. This paper addresses the particular reality of a non-profit organization, the Portuguese Lung Foundation (*Fundação Portuguesa do Pulmão* - FPP), that carry out several activities to collect, store and analyze data related with several diseases. The obtained results are used to characterize the actual reality and to set up campaigns aiming to improve the citizens' quality of life.

In particular, this paper is focused on the Chronic Obstructive Pulmonary Disease (COPD) for which Business Intelligence concepts and technologies are used to store and analyze the related data. The COPD is an airflow limitation that is not fully reversible and that affects up to one quarter of the adults with 40 or more years [1]. This disease is characterized by some of the following symptoms: chronic cough, sputum production and dyspnea. It can be confirmed in a clinical exam called spirometry, if the obtained values are 80% below of the Forced Expiratory Volume in 1 second (FEV1) and the ratio FEV1/FVC (Forced Vital Capacity) is lower than 0,7 [2]. The risk factors usually include: masculine gender, tobacco smoke, exposure to dusts and chemicals, air pollution, asthma, and genetic factors as a rare hereditary deficiency of α_1-antitrypsin [2]. This disease can be classified in four stages, according to the degree of severity. The first stage, or Mild COPD, is characterized by a FEV1 value above or equal to 80% and the presence, or not, of chronic cough and sputum production. COPD is not usually detected at this first stage. The second stage, or Moderated COPD, is characterized by a value of the FEV1 between 50% and 79%, shortness of breath during exertion, chronic cough and sputum production. The third stage, or Severe COPD, is characterized by a value of the FEV1 between 30% and 49%, greater shortness of breath, reduced exercise capacity and fatigue. The fourth stage, or Very Severe COPD, is characterized by a value of FEV1 below 30% and the presence of chronic respiratory failure [1] [2].

The analysis of the incidence of COPD and its geographical characterization is needed in order to provide health specialists with decision support indicators. To accomplish this goal, this paper presents the analysis of a data set made available by the FPP with data collected during 2007. The objective of this work is to design and implement a business intelligence system, analyze the data using On-Line Analytical Processing (OLAP) technology, and apply a spatial data mining algorithm for the identification of the spatial incidence and distribution of COPD. For the identified patterns, a characterization of them will be carried out in order to better understand, treat and prevent this disease. With the proposal of this business intelligence system, an integrated environment for the collection, storage and analysis of data is made available to the FPP. To the best of our knowledge, no similar system has been proposed and implemented. This business intelligence system includes a data mart model that enhances the data analysis tasks.

The advantage of applying business intelligence systems on real cases is associated to the delivery of innovative structures aimed to solve real world problems, giving a contribution to the theory of the area. This kind of approach has already been applied. We can find examples in the study of the toxic vigilance in France [3] and the study of sharing adverse drug event data [4].

This paper is organized as follows. Section II presents a brief overview on the available data and the transformations carried out to clean and put the data in the proper format for analysis. Section III describes the architecture of the proposed business intelligence system and the data mart

 ISBN: 978-1-61208-160-1

model used to store the data. Section IV presents the results obtained analyzing the data with OLAP technology. Section V describes the clustering approach and the SNN (Shared Nearest Neighbor) algorithm used to spatially characterize the incidence of COPD. Section VI concludes with some remarks and guidelines for future work.

II. AVAILABLE DATA

The data set available for this study was collected by the FPP in initiatives undertaken in Portugal in 2007. These initiatives are open to everyone who wants to participate. In them, the participants are asked to answer a questionnaire that integrates questions related to the symptoms and the risk factors of the COPD. The questionnaire also includes information about the geographical location where patients live (in a qualitative form), and their gender, age, height and weight. The result of the spirometry exam is also recorded.

The collected data, 1880 records, were made available in an Excel file. After an extensible analysis of the data, missing data fields and errors in data were identified.

Some of the tasks needed to clean (or to prepare) the data set include: i) the labeling of records without geographical localization (these records cannot be used in the spatial data mining task); ii) the replacing of *null* values on the height and weight with the mode of each one of these attributes; and, iii) the filling of the *do not know* stamp in all categorical attributes containing *null* values.

After the data cleaning process, a transformation phase took place. It was necessary to add the coordinates (x, y) of the geographical locations to each record. This will allow the use of the spatial data mining algorithm taking into consideration the geographical positions of the patients (the places where they live).

In the OLAP analysis presented afterwards, a subset of the available data is used. Only those records (275) related to patients with a FEV1 value lower than 80% are considered. This allows the characterization of the symptoms of individuals with COPD. For the spatial data mining task two different approaches are used. In the first, the whole data set is used, providing an overall characterization of the available data. After that, the subset with the 275 records, as in the OLAP analysis, is used.

For confidentiality reasons, no details about the available data are provided. Only aggregated results, resulting from OLAP analysis and the data mining process, are provided in Sections IV and V.

III. BUSINESS INTELLIGENCE SYSTEM

Business Intelligence systems are defined as analytical tools that aim the analysis of organizational data to provide further information to managers, improving the decision making process [8]. The Business Intelligence (BI) concept emerged as an evolution of the Decision Support Systems (DSS) [9]. This new concept replaced the data-oriented DSS because many approaches developed to assist the decision making process include OLAP and Data Mining technologies. OLAP arises to overcome some of the difficulties in the analysis of data stored in Operational Databases (ODB), which are exposed to continuous updates. The analytical process should access another database, specifically designed to support these analyses, the Data Warehouse (DW). It is on the data stored in this repository that OLAP and Data Mining technologies are usually applied in a BI context.

In this study, a BI system was designed and implemented to store and analyze the data collected by the FPP. The ETL (Extract, Transform and Load) process will take as input the data available in a spreadsheet and will run the appropriate mechanisms to clean, transform and load the data into the proposed data mart. This data mart will support the OLAP technology and the data mining algorithms.

To support forthcoming initiatives of the FPP, a web application dedicated to data collection tasks was implemented. The data collected through this web application is stored in an ODB enhancing the ETL process and improving the quality of the collected data.

The architecture of the BI system envisaged for the FPP is illustrated in Figure 1.

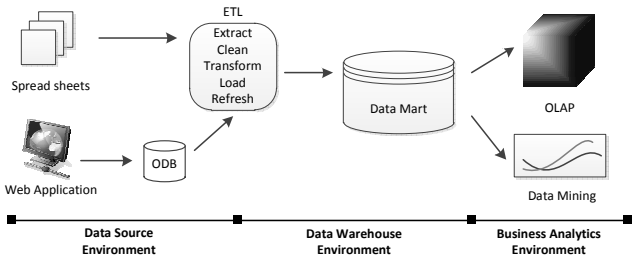

Figure 1. BI System Architecture

Related to implementation, the web application was developed using client/server pages (.ASP) and the programming languages HTML and Java Script. The ODB, the ETL process and the OLAP cubes were implemented using SQL Server 2008© technology. The Spatial Data Mining component makes use of a clustering algorithm implemented in Visual Basic© (more details can be found in Section V).

The Data Mart model was defined with the decision process in mind. This data model integrates one vector of analysis, represented by the fact table **FactFPP**. The star schema is shown in Figure 2. The **FactFPP** fact table allows the storage of the relevant information collected with the questionnaire. This fact table is linked to a set of dimension tables allowing the analysis of the available data in different perspectives. If we look to the data mart model, we can see that this table is linked to the **Time**, **Location**, **Profession**, **Patient**, **Smoke Characterization**, **Allergy Characterization**, **Cough Characterization**, **Fatigue Characterization** and **Pulmonary Diseases Characterization** dimension tables, meaning that the **FVC**, **FEV1**, **FEF 25-75** (Forced Expiratory Flow 25%–75%), three values obtained during the spirometry exam that characterize COPD, and the **Severity Stage**, can be analyzed by when, where the COPD is verified, who (with the information of the associated individuals such as age, gender, weight, among other attributes) and how (with the several questions of the questionnaire grouped in five distinct dimensions). The fact **Patient** is an event counter used to quantify the

number of individuals with specific symptoms or characteristics. It should be noted that the dimension **Profession** and the attributes marked with * in Figure 2 (attributes present in other dimensions or even in the fact table) will not be used in this first analysis (both OLAP and data mining) because these data are not present in the available data set (they will be included in future analysis when the web application is used for data collection).

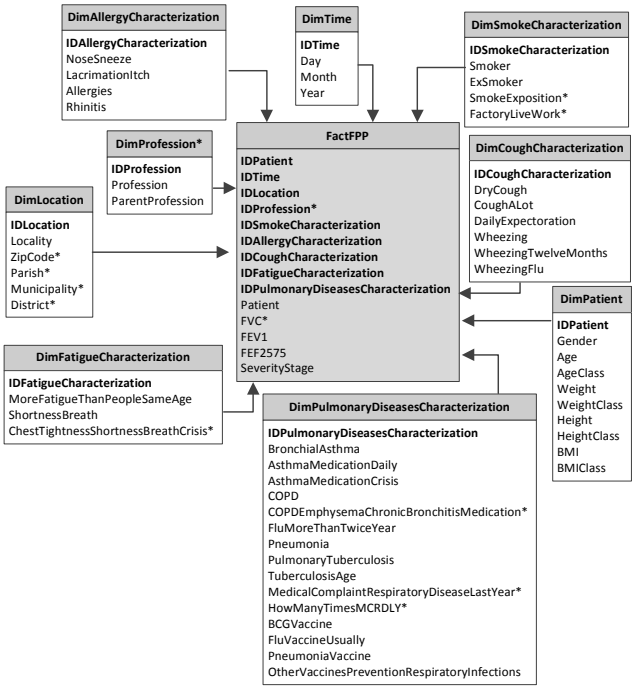

Figure 2. The Data Mart model

The presented data mart was designed in such a way that the star schema can evolve to a constellation schema as new fact tables are added. At this moment, two additional fact tables are envisaged: one for the study of pneumonia and the other for the study of lung cancer. As the constellation grows, more symptoms and data about the individuals can be related in the study of one or more diseases.

IV. DATA ANALYSIS WITH OLAP

After the presentation of the proposed BI system and the data model that stores all these data, this section presents the results obtained from the analysis of the available data using the OLAP technology. This technology is used to analyze the fact table along different perspectives.

The first analysis verifies the COPD severity stage of the individuals. Figure 3 shows that almost all the patients, 254 (92.7%), are at the 2nd COPD severity stage (Moderate). Only 18 patients are at the 3rd COPD severity stage (Severe) and 3 patients are at the 4th COPD severity stage (Very Severe).

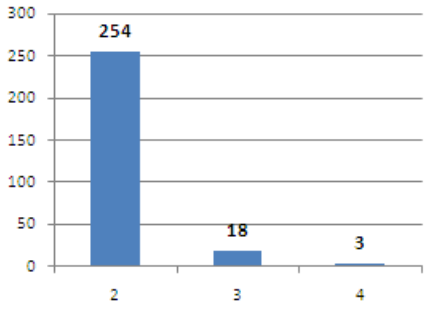

Figure 3. Severity Stages

After analyzing the COPD severity stages of the patients, the next analyses are focused on the answers given by the patients to the questions that allow their characterization. As already mentioned, all these individuals have a diagnosis of COPD after the spirometry exam. The characterization obtained to the group of questions related to **Smoke** can be seen in Figure 4. The results show that despite the tobacco is one of the most obvious risk factors for this disease, on the analyzed data, 168 patients (61.1%) with a diagnosis of COPD revealed that they never smoked.

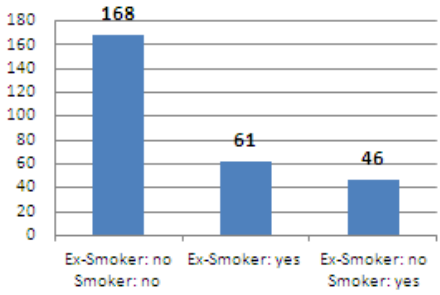

Figure 4. Smoke Characterization

Concerning the characterization of the group of questions **Fatigue**, the results are presented in Figure 5. Although it is clear that most of the patients, 187 (68.0%), feel more fatigue than people who have the same age (**MFTPSA**) and/or have shortness of breath (**SB**), and 118 of them (42.9%) even feel both symptoms, there are 88 individuals (32.0%) who have COPD that do not present any of these two symptoms.

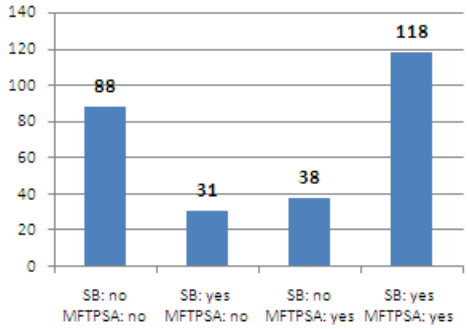

Figure 5. Fatigue Characterization

Analyzing the answers obtained to the group of questions **Cough** (Figure 6), and more precisely the two symptoms of COPD present in this group of questions, cough (**Dry Cough**) and expectoration (**Daily Expectoration – DE**), we can verify that 171 (62.2%) individuals with COPD have dry cough and daily expectoration. These two symptoms seem to be related. Indeed, when the individuals do not have dry cough, only 17.3% of them have daily expectoration. However, when the individuals have dry cough, the percentage of them who also have daily expectoration increases considerably (reaching 42.7%).

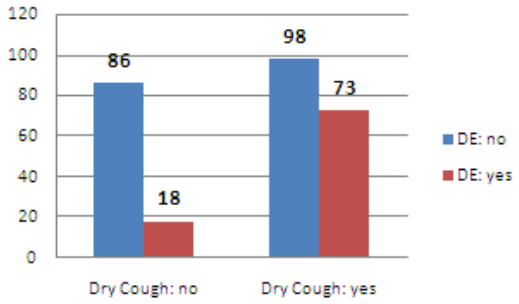

Figure 6. Cough Characterization

Concerning the **Pulmonary Diseases** group (TABLE I)**,** we analyzed the three pulmonary diseases characterized in this group of questions, which could have any relation with COPD. These diseases are the following: bronchial asthma (a COPD risk factor), pneumonia and pulmonary tuberculosis (**Pulm Tuber**).

Analyzing the cube at TABLE I, none of the individuals suffered from all these three diseases (there are cases of **dk** – *do not know* – that are not considered in this observation). Another observation that can be pointed from TABLE I is that only a low percentage of these patients who have COPD suffered from at least one of these three diseases. Indeed, 65 (23.6%) individuals have or had bronchial asthma, 57 (20.7%) individuals suffer or suffered from pneumonia and 17 (6.2%) from pulmonary tuberculosis. As bronchial asthma is a risk factor of COPD, it could be expected that the percentage of incidence of this disease was significantly higher in patients who have COPD.

TABLE I. PULMONARY DISEASES CHARACTERIZATION

		Bronchial Asthma		
		no	yes	Grand Total
Pneumonia	Pulm Tuber	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ dk	dk	3	2	5
	no	5	2	7
	Total	8	4	12
⊟ no	dk	2	2	4
	no	155	35	190
	yes	9	3	12
	Total	166	40	206
⊟ yes	dk	6	1	7
	no	25	20	45
	yes	5		5
	Total	36	21	57
Grand Total		210	65	275

Regarding the answers obtained to the group of questions **Allergy**, the results are presented in Figure 7. We analyzed the incidence of three characteristics from this group of questions: the eyes lacrimation and itch (**LI**), runny nose and sneeze (**NS**), and rhinitis. The results showed that 202 (73.5%) individuals with COPD also have runny nose and sneeze when they are not with flu. This fact can reveal a link between these symptoms and COPD. However, there is almost the same number of patients, 201 (73.1%), who claims that they do not suffer from rhinitis. Another fact observed in this group of questions was that from the 73 patients with COPD who do not have runny nose and sneeze, 10 (13.7%) have eyes lacrimation and itch. From the 202 patients with COPD who have runny nose and sneeze, 128 (63.4%) have eyes lacrimation and itch.

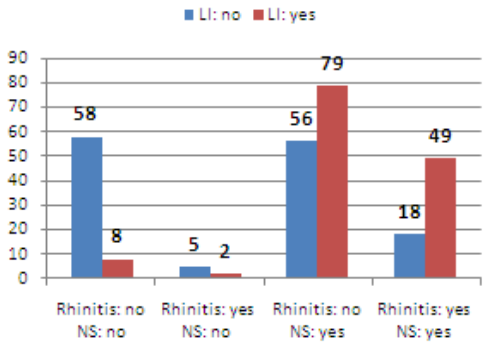

Figure 7. Allergy Characterization

This last data analysis with OLAP has as objective to show the importance of the initiatives and campaigns made in favor of this respiratory disease and the need to participate in them. TABLE II shows the obtained results. Patients were subdivided by **Gender** and by **Age Class**. The **COPD** attribute indicates if the patient already knew that he/she suffers from COPD before the spirometry exam. We can say, based on these data, that 229 patients (83.3%) with COPD did not know that they had this disease before participating in these initiatives promoted by the FPP. This cube also shows age as a risk factor for COPD. Indeed, 204 individuals (74.2%) with more than 41 years old have this disease. It was also noted that only 7 (2.5%) individuals, between 0 and 17 years old, have COPD.

TABLE II. IDENTIFICATION OF COPD AT THE FPP'S INITIAVIVES

		COPD		
		no	yes	Grand Total
Gender	Age Class	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ f	[0-17]	3		3
	[18-40]	37	2	39
	[41-64]	54	8	62
	65+	44	9	53
	Total	138	19	157
⊟ m	[0-17]	4		4
	[18-40]	22	3	25
	[41-64]	29	4	33
	65+	36	20	56
	Total	91	27	118
Grand Total		229	46	275

 ISBN: 978-1-61208-160-1

V. DATA ANALYSIS WITH SPATIAL DATA MINING

This section presents the analysis of the available data using a data mining algorithm, given particular attention to the spatial component of the data. Firstly, it will be presented the data mining algorithm that will be used and then the obtained results.

A. The Shared Nearest Neighbor (SNN) Algorithm

The SNN algorithm was first presented by Levent Ertöz, Michael Steinbach and Vipin Kumar in 2003 on the "SIAM International Conference on Data Mining". This algorithm can find clusters with different sizes, shapes and densities, and automatically identifies the number of clusters in the data set [10]. Usually, this algorithm has three input parameters, *k* – the size of the list of neighbors, *Eps* – the density threshold, and the *MinPts* – the threshold value used to classify the core points [10] (Figure 8).

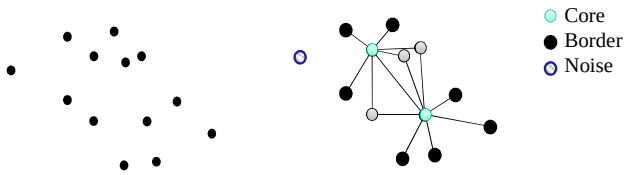

Figure 8. SNN: an example

The implementation used in this paper was coded from scratch in Visual Basic. It requires only an input parameter: *k*. *Eps* and *MinPts* are calculated based on *k* as follows: *Eps = 3k/10* and *MinPts = 7k/10* [11] and includes the following steps [11]:

1. Identify the "*k*" nearest neighbors for each point (using a distance function to calculate similarity);
2. Calculate the SNN similarity between pairs of points as the number of nearest neighbors that the two points share;
3. Calculate the SNN density of each point;
4. Detect the core points;
5. Form the clusters from the core points;
6. Identify the noise points;
7. Assign the remainder points to the cluster that contains the most similar core point.

In the work presented in this paper were used two different distance functions: one for all the available data and another for records with FEV1<80%. This way it is possible to mine the available data and find spatial distributions taking into consideration the whole data set, patients with or without COPD, and also those records that are exclusively associated with patients that suffer from COPD. For this last data set, it is possible to group patients not only by their geographical proximity but also by their incidence of COPD. The equations to the distance functions are the following (adapted from [6]):

$$DistFunction(p_1, p_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (1)$$

$$DistFunction(p_1, p_2) = w_1 * \left(\frac{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}}{mDist} \right) + w_2 * \left(\frac{|s_1 - s_2|}{mFEV} \right) \quad (2)$$

In equation 2, w_1 and w_2 represent the weights assigned to the position and to the FEV1 parameters. Through these weights it is possible to verify the impact of the position and the FEV1 value on the clustering result. Also, in equation 2, *mDist* and *mFEV* are the maximum difference values between any two points (for the distance and for the FEV1, respectively). They are used to normalize the distance function value.

In order to identify the appropriate values for w_1 and w_2, some simulations were made with the data set. The obtained results are presented in Figure 9.

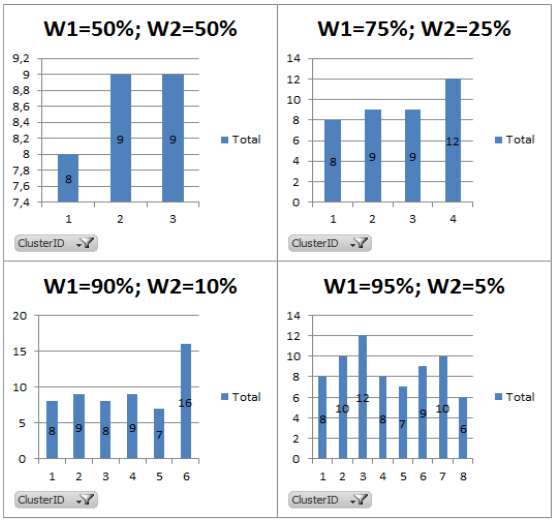

Figure 9. W_1 and W_2 analysis

After analyzing the obtained results, the weights w_1=95% and w_2=5% seem to be more appropriate as they allow the identification of more clusters, with different shapes and sizes. The other weights classify a large number of records as noise. These weights already showed to be appropriate in other contexts [6].

B. Obtained Results

As mentioned in Section II, we are using two different approaches. In the first, all the records available in the data set are used. After several simulations to identify a proper value for *k*, we decided to adopt a *k* value of 15, which returned 41 clusters, with the spatial distribution shown in Figure 10. In this figure each cluster is represented by a different color. As the coordinates were added at the Parish level, and not at a lower level of detail, a small agitation of the points coordinates was carried out in order to avoid points overlapping in the clusters visualization.

From the presented results it is possible to verify that the majority of the points are associated to the huge metropolitan areas of Oporto and Lisbon, despite the big clusters (in area) spread at the interior, central coast and south of the country. These results can have two main explanations: the fact that these two regions are the most populated in Portugal or a non-balanced data set with a higher incidence of individuals from these two metropolitan areas. This last case was the verified in the data set under analysis.

 ISBN: 978-1-61208-160-1

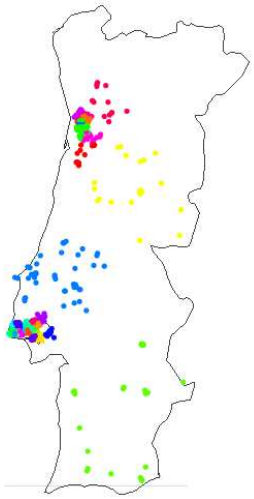

Figure 10. Clusters for the first approach to the data analysis

Associated with the presented results, we proceed verifying the risk factors associated to the patients of this first approach. One of the risk factors is easy to see just by looking at the graphical distribution of the points, the air pollution, because the majority of the clusters are located in the more populated regions of the country and where the incidence of factories, cars and other pollutants is higher. To analyze the other risk factors, the answers to the questionnaire were verified (Figure 11). Analyzing the risk factors we verify that 44% of the patients that are grouped in the several clusters are or were smokers, 45% are male, 17% are older than 65 years old, and 15% suffers from asthma.

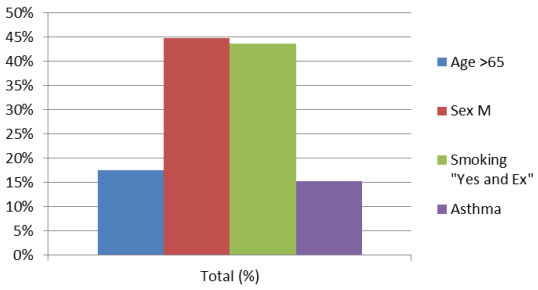

Figure 11. % total of risk factors (first set of results)

The other approach followed only uses the records that have a FEV1 value below 80%. In this second approach the equation of the distance function used is the equation 2 and after some simulations the best value found for *k* is 7. The obtained results are presented in Figure 12 where colors are associated to clusters (each cluster has a different color).

Comparing Figure 12 with Figure 10 it is possible to verify that the clusters in Figure 12 are more concentrated in one region which is the region of Lisbon and surroundings. The other cluster appears at North, grouping individuals from several districts. Obviously, more data is needed to obtain a deeper understating on the spatial distribution of COPD.

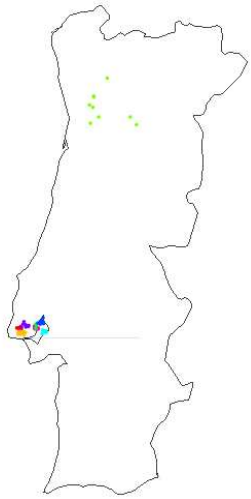

Figure 12. Clusters for the second approach to the data analysis

In a more detailed analysis to this second set of results is possible to see that, as in the first set of results, the more prominent risk factors are being male, smoking or advanced age (Figure 13), with exception of: cluster 2 where the asthma has more incidence than smoking; and, cluster 4 with asthma having the same incidence as male, smoking and advanced age.

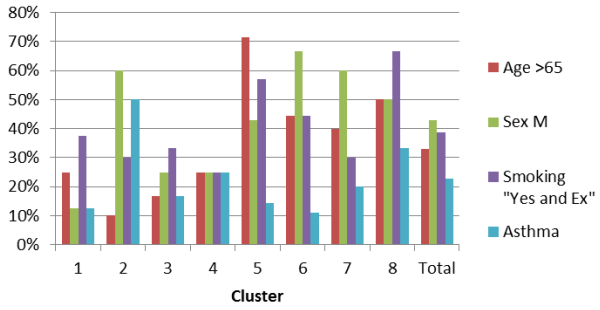

Figure 13. % Risk factors per cluster and total

Looking at the symptoms (Figure 14), the two symptoms that appear with more incidence are the fatigue and the shortness of breath, exception made to the first cluster where the chronic cough is the symptom that appear in first place. This incidence of fatigue and shortness of breath has huge impact on the quality of life of the patients. The fact that these patients are from the more populated regions in Portugal, and consequently with more pollution, and nearly 40% of those are smokers or ex-smokers, could help to explain why these two symptoms have the higher percentages. Since the answers to the questionnaire do not tell us if patients have chronic respiratory failure and the values of FEV1 are all above 50%, the stage of COPD in which the patients are is the stage 2 or moderate COPD (TABLE III). All the patients with the FEV1 below 50% were considered as noise by the clustering algorithm because they are only a few records, and the distribution of the points in terms of geographic locations and the FEV1 values do not

 ISBN: 978-1-61208-160-1

have sufficient similarities to group them in one or more clusters.

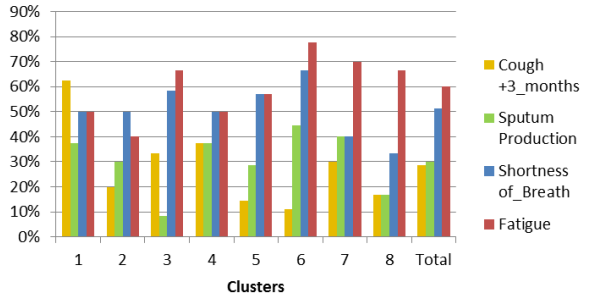

Figure 14. % Symptoms per cluster and total

TABLE III. COPD STAGES PER CLUSTER

Stages COPD	**Cluster**							
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Stage 2	8	10	12	8	7	9	10	6
Stage 3	0	0	0	0	0	0	0	0
Stage 4	0	0	0	0	0	0	0	0
Total	**8**	**10**	**12**	**8**	**7**	**9**	**10**	**6**

VI. CONCLUSION

This paper presented a business intelligence system for the analysis of data collected by the Portuguese Lung Foundation (FPP – *Fundação Portuguesa do Pulmão)*. The objectives set to this work were to design and implement a business intelligence system, analyze the data using OLAP, and apply a spatial data mining algorithm for the identification of the spatial incidence and distribution of COPD (Chronic Obstructive Pulmonary Disease). The implemented system integrated a data mart for the storage of data and OLAP and spatial data mining technologies for data analysis.

Data from 1880 patients were available from the FPP´s data set. Each patient answered a questionnaire and made a clinical exam called spirometry. The several analyses in terms of OLAP reinforced essentially the idea that this disease is very difficult to diagnose without the spirometry exam. That happens because despite we have confirmed some risk factors related to COPD and identified some patterns of this disease, there seems to be a kind of contradiction with their symptoms. This happens, for instance, with the characterization of the responses obtained in the question groups **Fatigue** and **Cough**. Many patients do not have any symptoms from these two groups of questions that are typical of this respiratory disease. Therefore, the difficulty in diagnosing the disease also reinforce the conclusion that a good prevention of COPD, making a spirometric exam periodically, is essential for people who fall within the risk factors of this disease.

In terms of spatial analysis we have two main conclusions. First, when comparing the data between the two set of results, the distribution of the clusters is similar and the percentages of the risk factors are also similar. The second conclusion is that the main symptoms that the patients mention to have are shortness of breath and fatigue, and all the patients grouped in the clusters are in the stage 2 of COPD.

The proposed BI system showed to be useful in the analysis of the available data, allowing the characterization of the key factors that are associated to COPD.

Related to future work, it is envisaged to perform the analysis with more data, to automate the ETL process between the operational database of the web application and the data mart, to incorporate a temporal variable tagging each patient in order to be possible a spatio-temporal analysis, and to perform analysis with other spatial cluster algorithms besides the SNN used in this paper.

REFERENCES

[1] GOLD, Global Strategy for Diagnosis, Management, and Prevention of COPD, Tecnical Report, Global Initiative for Chronic Obstructive Lung Disease, 2010.

[2] R. A. Pauwels, A. S. Buist, P. M. A. Calverley, C. R. Jenkins, and S. S. Hurd, “Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease,” American Journal of Respiratory and Critical Care Medicine, vol. 163, pp. 1256-1276, 2001.

[3] G. Guyodo, I. Blanc, J. L. Boulben, Lefevre B, F. De Bels, and R. Garnier, “The French Toxic Exposure Surveillance System: Adaptation of a Business Intelligence System for Toxicovigilance,” Clinical Toxicology, vol. 48, no. 3, 2010.

[4] M. Horvath, H. Cozart, A. Ahmad, M. K. Langman, and J. Ferranti, “Sharing Adverse Drug Event Data Using Business Intelligence Technology,” Journal of Patient Safety, vol. 5, no. 1, pp. 35-41, 2009.

[5] S. Brooker, S. Clarke, J. K. Njagi, S. Polack, B. Mugo, B. Estambale, E. Muchiri, P. Magnussen, and J. Cox, “Spatial Clustering of Malaria and Associated Risk Factors During an Epidemic in a Highland Area of Western Kenya,” Tropical Medicine & International Health, vol. 9, no. 7, pp. 757-766, 2004.

[6] A. Moreira, M. Y. Santos, M. Wachowicz, and D. Orellana, "The Impact of Data Quality in the Context of Pedestrian Movement Analysis," In M. Painho, M. Y. Santos and H. Pundt (Eds.), Geospatial Thinking, Lecture Notes in Geoinformation and Cartography: Springer, 2010.

[7] G. R. Jordan, N. Loveridge, K. L. Bell, J. Power, N. Rushton, and J. Reeve, “Spatial clustering of remodeling osteons in the femoral neck cortex: a cause of weakness in hip fracture?,” BONE, vol. 26, no. 3, pp. 305-313, 2000.

[8] W. F. Cody, J. T. Kreulen, V. Krishna, and W. S. Spangle, “The integration of business intelligence and knowledge management,” IBM Systems Journal vol. 41, pp. 697-713, 2002.

[9] S. Alter, Information Systems: A Management Perspective: Addison Wesley Longman, 1999.

[10] L. Ertöz, M. Steinbach, and V. Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data", Proceedings of the Second SIAM International Conference on Data Mining, San Francisco, pp. 47-58, 2003.

[11] A. Moreira, M. Y. Santos, and S. Carneiro, "Density-based clustering algorithms – DBSCAN and SNN (Implementation Documentation)," http://ubicomp.algoritmi.uminho.pt/local/, Date of access: February, 2011.

 ISBN: 978-1-61208-160-1

Enforcing the Repeated Execution of Logic in Workflows

Mirko Sonntag, Dimka Karastoyanova
Institute of Architecture of Application Systems
University of Stuttgart, Universitaetsstrasse 38
70569 Stuttgart, Germany
{sonntag, karastoyanova}@iaas.uni-stuttgart.de

***Abstract*—The repeated execution of workflow logic is a feature needed in many situations. Repetition of activities can be modeled with workflow constructs (e.g., loops) or external workflow configurations, or can be triggered by a user action during workflow execution. While the first two options are state of the art in the workflow technology, the latter is currently insufficiently addressed in literature and practice. We argue that a manually triggered rerun operation enables both business users and scientists to react to unforeseen problems and thus improves workflow robustness, allows scientists steering the convergence of scientific results, and facilitates an explorative workflow development as required in scientific workflows. In this paper, we therefore formalize operations for the repeated enactment of activities—for both iteration and re-execution. Starting point of the rerun is an arbitrary, manually selected activity. Since we define the operations on a meta-model level, they can be implemented for different workflow languages and engines.**

Keywords-service composition; workflow adaptability; iteration; re-execution.

I. INTRODUCTION

Imperative workflow languages are used to describe all possible paths through a process. On the one hand, this ensures the exact execution of the modeled behavior without deviations. On the other hand, it is difficult, if not impossible, to react to unforeseeable and thus un-modeled situations that might happen during workflow execution, e.g., exceptions. This is the reason why flexibility features of workflows were identified as essential for the success of the technology in real world scenarios (e.g., [1]). In [2], four possible modifications of running workflows are described as advanced functions of workflow systems: the deletion of steps, the insertion of intermediary steps, the inquiry of additional information, the iteration of steps.

In this paper, we focus on the iteration of steps. Usually, iterations are explicitly modeled with loop constructs. But not all eventualities can be accounted for in a process model prior to runtime. Imagine a process with an activity to invoke a service. At runtime, the service may become unavailable. The activity and hence the process will fail, leading to a loss of time and data. Rerunning the activity (maybe with modified input parameters) could prevent this situation.

The repetition of workflow logic is not only meaningful for handling faults. In the area of scientific workflows, the result of scientific experiments or simulations is not always known a priori [3, 4]. Scientists may need to take adaptive actions during workflow execution. In this context, rerunning activities is basically useful to enforce the convergence of results, e.g., redo the generation of a Finite Element Method (FEM) grid to refine a certain area, repeat the visualization of results to obtain an image with focus on another aspect of a simulated object, or enforce the execution of an additional simulation time step.

A lot of work exists on the repetition of activities in workflows. Existing approaches use modeling constructs (e.g., BPEL retry scopes [5]), workflow configurations (e.g., Oracle BPM [6]), or automatically selected iteration start points (e.g., Pegasus [7]) to realize the repeated execution of workflow parts. An approach for the repetition of workflow logic with an arbitrary starting point that was manually selected at runtime by the user/scientist is currently missing. We argue that such functionality is useful in both business and scientific workflows. In business workflows it can help to address faulty situations, especially those where a rerun of a single faulted activity (usually a service invocation) is insufficient. In scientific workflows it is one missing puzzle piece to enable explorative workflow development [4, 8] and to control and steer the convergence of results.

In this paper, we therefore formalize two operations on workflow instances to enforce the repetition of workflow logic, namely iteration and re-execution. The *iteration* works like a loop that reruns a number of activities. The *re-execution* undoes work completed by a set of activities with the help of compensation techniques prior to the repetition of the same activities. We define the operations on the level of the workflow meta-model. Thus, the operations can be implemented in different workflow languages and engines. We discuss several problems that arise when repeating arbitrary workflow logic, such as data handling issues and the communication with clients.

The rest of the paper is organized as follows. Section II presents other work on the topic. Section III shows the workflow meta-model used in this work. Section IV describes the *iterate* and *re-execute* operations and discusses implications of the approach. Section V concludes the paper.

II. RELATED WORK

Repetition of workflow logic can be achieved language-based with certain modeling constructs. A general concept to retry and rerun transaction scopes in case of an error is shown in [9]. Eberle et al. [5] apply this concept to BPEL

 ISBN: 978-1-61208-160-1

scopes. In BPMN [10] this behavior can be modeled with sub-processes, error triggers and links. In $ADEPT_{flex}$ it is possible to model backward links to repeat faulted workflow logic [11]. In IBM MQSeries Workflow and its Flow Definition Language (FDL) activities are restarted if their exit condition evaluates to false. ADOME [12] can rerun special repeatable activities if an error occurs during activity execution. In Apache ODE an extension of BPEL invoke activities enables to retry a service invocation if a failure happens [13]. These approaches have in common that special modeling constructs realize the repetition. Thus, the iteration is pre-modeled at design time. Further, FDL and ADOME allow the rerun of a single faulted activity only. In contrast to these approaches, our solution aims at repeating a workflow starting from an arbitrary, not previously modeled point.

Iterations can also be realized by configuring workflow models with deployment information. Invoke activities in the Oracle BPEL Process Manager [6] can be configured with an external file so that service invocations are retried if a specified error occurs. The concept to retry activities until they succeed is also subject of [14]. The scientific workflow system Taverna [15] allows specifying alternate services that are taken if an activity for a service invocation fails. In contrast to these approaches, we advocate a solution where the rerun can be started spontaneously without a pre-configuration of workflows.

The scientific workflow system Pegasus can automatically re-schedule a part of a workflow if an error occurs [7]. Successfully completed tasks are not retried. Kepler's Smart Rerun Manager can be used to re-execute complete workflows [16]. Tasks that produce data that already exists are omitted. The main difference of these approaches to our idea is that we select the starting point of the iteration manually and hence can use this functionality for explorative workflow development and steering of the convergence of scientific results.

In [2] the iteration of activities is mentioned but without going into details. In the scientific workflow system e-BioFlow scientists can re-execute manually selected tasks with the help of an ad hoc workflow editor [3]. The set of activities that should be (re-)executed must be marked explicitly. No other activities are (re-)executed; no distinction is made between iteration and re-execution operations. In our approach the user only has to provide the start activity for the rerun and the successor activities are then executed as prescribed by the workflow model.

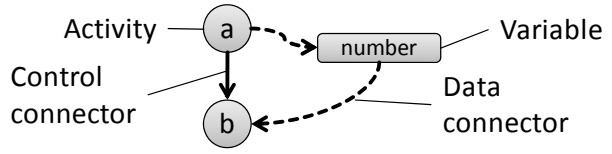

Figure 1. Example for a process model

III. META-MODEL

At first, we introduce main concepts of the workflow meta-model we use in this paper. We focus on those aspects of the meta-model needed to describe the repeated execution of workflow logic. A process model is considered a directed and acyclic graph (Figure 1). The nodes are tasks to be performed (i.e., activities). The edges are control connectors (links) and prescribe the execution order of activities. Data dependencies are represented by variables that are read and written by activities.

Definition 1 (Variable, V). The set of variables $V \subseteq M \times S$ (M = set of names; S = set of data structures) defines all variables of a process model [2]. Each $v \in V$ has assigned a finite set of possible values, its domain DOM(v) [2].

Definition 2 (Activity, A). Activities are functions that perform specific tasks. A join condition $j \in C$ with C as the set of all conditions can be assigned to an activity. If j evaluates to true at runtime, the activity is scheduled. The set of all activities of a process model is $A \subseteq M \times C$. Variables can be assigned to activities with the help of an input variable map $\iota: A \mapsto \wp(V)$ and an output variable map $o: A \mapsto \wp(V)$. Input variables may provide data to activities and activities may write data into output variables. Further, compensating activities that undo the effects of an activity can be assigned by a compensate activity map $c: N \mapsto N$.

Definition 3 (Link, L). The set $L \subseteq A \times A \times C$ denotes all control connectors in a process model. Each link connects a source with a target activity. Its transition condition $t \in C$ says if it is followed at runtime. Two activities can be connected with at most one link (i.e., links are unique).

Definition 4 (Process Model, G). A process model is a directed acyclic graph denoted by a tuple G = (m, V, A, L) with a name $m \in M$.

A. Execution And Navigation

For the execution of a process model, a new process instance of that model is created, activities are scheduled and performed, links are evaluated, and variables are read and written. These tasks (i.e., the navigation) are conducted according to certain rules. The component of a workflow system that supervises workflow execution and that implements these rules is called the navigator. We reflect the notion of time in the meta-model with ascending natural numbers. Each process instance possesses its own timeline. At time $0 \in \mathbb{N}$ a process is instantiated. Each navigation step increases the time by 1. Hence, navigation steps conducted in parallel have different time steps. In the following we present navigation rules that are most important for this work.

If an activity is executed, an activity instance is created with a new unique id. If the same activity is executed again (e.g., because it belongs to a loop), another instance of it is created with another id. The same holds for links and link instances. A new id can be generated with the function `newId()` that delivers an element of the set of ids, ID.

 ISBN: 978-1-61208-160-1

We consider process, activity and link instances sets of tuples. This allows us to navigate through a process by using set operations. Navigation steps are conducted by creating new tuples and adding them to sets (instantiation of an activity/a link) or by taking tuples from sets and adding modified tuples to sets (to change the state of existing activity/link instances).

Definition 5 (Variable Instance, $\mathcal{V}$). Variable instances provide a concrete value c for a variable v (i.e., an element of its domain) at a point in time t. The finite set of variable instances is denoted as $\mathcal{V} = \{(v, c, t) \mid v \in V, c \in DOM(v), t \in \mathbb{N}\}$. The set of all possible variable instances is $\mathcal{V}_{all}$.

Definition 6 (Activity Instance, $\mathcal{A}$). Each activity can be instantiated several times. These different instances are referred to by ids that are unique to activity instances. The set of activity instances is denoted as $\mathcal{A} = \{(id, a, s, t) \mid id \in ID, a \in A, s \in S, t \in \mathbb{N}\}$. At a point in time t an activity instance has an execution state $s \in S$ = {scheduled, executing, completed, faulted, terminated, compensated}. Note that an activity instance *a* reaches the compensated state if it is completed and its compensation activity $c(\pi_2(a))$ was executed successfully.

We define three sets that help to capture the state of a process instance and that are used to navigate through a process model graph.

Definition 7 (Active Activities, $\mathcal{R}$). The finite set of active activities $\mathcal{R}$ contains all activity instances that are scheduled or currently being executed: $\mathcal{R} \subseteq \mathcal{A}$, $\forall a \in \mathcal{R}$: $\pi_3(a) \in$ {scheduled, executing}.

Definition 8 (Finished Activities, $\mathcal{F}$). The finite set of finished activities $\mathcal{F}$ contains all activity instances that are completed, faulted or terminated: $\mathcal{F} \subseteq \mathcal{A}$, $\forall a \in \mathcal{F}$: $\pi_3(a) \in$ {completed, faulted, terminated}. Note that compensated activities are not part of $\mathcal{F}$ because their effects are undone.

Definition 9 (Active Links, $\mathcal{L}$). The finite set of active links $\mathcal{L}$ contains link instances that refer to the instantiated link e and a truth value for the evaluated transition condition: $\mathcal{L} = \{(e, t) \mid e \in E, t \in \{true, false\}\}$. $\mathcal{L}$ contains only those link instances that are evaluated but where the target activity is not yet scheduled or being executed: $\forall l \in \mathcal{L}$: $\nexists a \in \mathcal{R}$: $\pi_2(\pi_1(l)) = \pi_2(a)$.

Definition 10 (Wavefront, $\mathcal{W}$). The set of all active activities and links in a process instance is called the wavefront $\mathcal{W} = \mathcal{R} \cup \mathcal{L}$.

Definition 11 (Process Instance, p_g). An instance for a process model g is now defined as a tuple $p_g = (\mathcal{V}, \mathcal{R}, \mathcal{F}, \mathcal{L})$. The set of all process instances is denoted as $\mathcal{P}_{all}$.

As navigation example consider Figure 1. Say activity $a \in A$ is currently being executed and invokes a program that increases a given number by 1. The process instance thus looks as follows: p_g = ({(number, 100, 1)}, {(382, a, executing, 3)}, {}, {}). If activity *a* completes, its corresponding tuple is deleted from $\mathcal{R}$ and a new tuple with the new activity instance state and an increased time step is added to $\mathcal{F}$: p_g = ({(number, 100, 1)}, {}, {(382, a, completed, 4)}, {}). Now, the navigator stores the new value of the variable *number* and deletes the former value: p_g = ({(number, 101, 5)}, {}, {(382, a, completed, 4)}, {}). Even though the navigator manipulates the tuples, all these actions are recorded in the audit trail.

IV. ITERATION AND RE-EXECUTION

Based on the meta-model described above we can now address the repeated execution of workflow parts. As already proposed in [5], we also want to distinguish between two repetition operations. The first operation reruns workflow parts without taking corrective actions or undoing already completed work. The second operation resets the workflow context and execution environment with compensation techniques prior to the rerun (e.g., de-allocating reserved computing resources, undoing completed work).

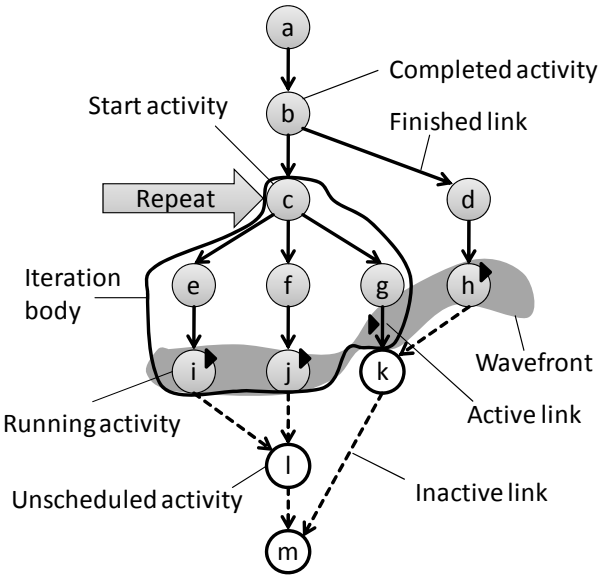

Figure 2. Example of a process instance

Before we dive into the details of the iteration of workflow parts we have to introduce several important terms (see Figure 2). The point from where a workflow part is executed repeatedly is denoted as the *start activity* (activity *c* in the figure). The start activity is chosen manually by the user/scientist at workflow runtime. The workflow logic from

the start activity to those active activities and active links that are reachable from the start activity are called *iteration body* (activities *c, e, f, g, i, j*, the links in between and link *g-k*). The iteration body is the logic that is executed repeatedly. Note that activities/links reachable from the iteration body but not in the iteration body are executed normally when the control flow reaches them (e.g., activities *k* and *l*).

For the iteration/re-execution of logic it is important to avoid race conditions, i.e., situations where two or more distinct executions of one and the same path are running in parallel. These situations can occur in cyclic workflow graphs or can be introduced by the manual rerun of activities that we propose. For example, if the repetition is started from activity *c* in Figure 2, then a race condition emerges because activities *i* and *j* on the same path are still running: activity *l* could be started if *i* and *j* complete while a competing run is started at *c*. There are two ways to avoid race conditions in this scenario. Firstly, the workflow system can wait until the running activities in the iteration body are finished without scheduling any successor activities (here: *l*). The rerun is triggered only afterwards. Secondly, running activities in the iteration body can be terminated and the rerun can start immediately. A workflow system should provide both options to the users because in some cases it is meaningful to complete running work prior to the rerun while in other cases the result of running work is unimportant. This has to be decided on a per-case-basis by the user. In the rest of this paper we focus on the more complex second option: termination. We therefore need to define an operation that terminates running activities.

Definition 12 (Termination). The terminate operation prematurely aborts running activities. Let $n = (id, a, s, t) \in \mathcal{R}$ with $s \in \{active, executing\}$ be an activity instance. Then `terminate(n)` delivers the tuple (id, a, terminated, t') with $t' > t$, i.e., the activity instance is terminated.

Definition 13 (Active Successor Activities). We need a function that finds all activities in the wavefront that belong to the iteration body. The function delivers exactly those running activities that have to be terminated before the rerun can be started:

$$\text{activeSuccActivities} : A \times \mathcal{P}_{all} \mapsto \wp(\mathcal{R})$$

Let $a \in A$ be an activity in process model g and $p_g \in \mathcal{P}_{all}$ an instance of g. Then `activeSuccActivities`$(a, p_g) = \{r_1, \ldots, r_k\}$, $r_1, \ldots, r_k \in \mathcal{R} \Leftrightarrow \forall i \in \{1, \ldots, k\}: \pi_2(r_i)$ is reachable from *a*.

Race conditions can also occur if active links remain in the process instance. In Figure 2, a race condition could appear as follows. If activity *k* completes and the link *h-k* is evaluated, the join condition of *k* could become true. *k* would then be started although a competing execution of the same path arises due to the repetition of *c*. That is why such links have to be found and reset.

Definition 14 (Active Successor Links). A function is needed that finds all links in the wavefront where the source activity is reachable from a given activity:

$$\text{activeSuccLinks}: A \times \mathcal{P}_{all} \mapsto \wp(\mathcal{L})$$

Let $a \in A$ be an activity of process model g and $p_g \in \mathcal{P}_{all}$ an instance of g. Then `activeSuccLinks`$(a, p_g) = \{l_1, \ldots, l_k\}$, $l_1, \ldots, l_k \in \mathcal{L} \Leftrightarrow \forall i \in \{1, \ldots, k\}: \pi_1(l_i)$ is reachable from *a*.

A. Iteration

Parts of a workflow may be repeated without the need to undo any formerly completed work. A scientist may want to enforce the convergence of experiment results and therefore repeats some steps of a scientific workflow. This is what we denote as iteration of workflow parts.

Definition 15 (Iterate Operation). The iteration is a function that repeats logic of a process model for a given process instance beginning with the given activity as starting point and taking the data indicated by the given time step as input for the next iteration.

$$\mathfrak{i} : A \times \mathcal{P}_{all} \times \mathbb{N} \mapsto \mathcal{P}_{all}$$

Let $a \in A$ be the start activity of the iteration and $p\_in_g, p\_out_g \in \mathcal{P}_{all}$ two process instances. Here, p_in_g is the input for the $\mathfrak{i}$ operation and p_out_g is the resulting instance with changed state that is ready to start with the iteration. As pre-condition we define that only already executed activities can be used as start activity: $\exists n \in \mathcal{R} \cup \mathcal{F} : \pi_2(n) = a$. This prevents (1) using the operation on dead paths, (2) jumping into the future of a process instance, and (3) guarantees the correct termination of running activities. Then $\mathfrak{i}(a, p\_in_g, t) = p\_out_g$ with $t \in \mathbb{N}$, $p\_in_g = (\mathcal{V}_{in}, \mathcal{R}_{in}, \mathcal{F}_{in}, \mathcal{L}_{in})$ and $p\_out_g = (\mathcal{V}_{out}, \mathcal{R}_{out}, \mathcal{F}_{out}, \mathcal{L}_{out}) : \Leftrightarrow$

1. $\mathcal{V}_{out} = \mathcal{V}_{in}$
2. $\mathcal{R}_{out} = \mathcal{R}_{in} \setminus \text{activeSuccActivities}(a) \cup \{($`newId()`$, a, \text{active}, i)\}$, *i* is a new and youngest time step
3. $\mathcal{F}_{out} = \mathcal{F}_{in} \cup \cup_{n \in \text{activeSuccActivities}(a)} \text{terminate}(n)$
4. $\mathcal{L}_2 = \mathcal{L}_1 \setminus \text{activeSuccLinks}(a)$

The variables remain unchanged (1.). All active successor activities from *a* are terminated, i.e., deleted from the set of running activities (2.) and inserted with a new status to the set of finished activities (3.). All active links in the iteration body are reset (4.). The start activity is scheduled (added to the set of active activities with status *active*) so that the workflow logic is repeated beginning with the start activity (2.). The join condition of the start activity is not evaluated again.

B. Re-execution

It is also needed to repeat parts of a workflow as if they were executed for the first time. Completed work in the iteration body has to be reversed/compensated prior to the repetition. A scientist may want to retry a part of an experiment because something went wrong. But the

 ISBN: 978-1-61208-160-1

execution environment has to be reset first. This is what we denote as re-execution of workflow parts.

Algorithm 1 (Compensate Iteration Body). For the compensation of completed work in the iteration body we propose an algorithm with the following signature:

compensateIterationBody : A $\times$ $\mathcal{P}_{all} \mapsto \wp(\mathcal{V}_{all})$

The function compensates all completed activities of the iteration body in reverse execution order. It delivers the values of variables that were changed during compensation. Let $a \in$ A be the start activity of the re-execution and $p \in \mathcal{P}_{all}$ a process instance for the model of *a*. Then compensateIterationBody(*a*, *p*) = {v_1, …, v_k} with $p = (\mathcal{V}, \mathcal{R}, \mathcal{F}, \mathcal{L})$, v_1, … , $v_k \in \mathcal{V}_{all}$ works as follows (Note: f.state for f $\in \mathcal{F}$ is equivalent to π_3(f); f.time to π_4(f)):

```
function compensateIterationBody(a, p)
1  V_result ← ∅
2  F = {f ∈ F | f.state == completed ∧
   π2(f) is reachable from a}
3  while (|F| > 0) do
4   if |F| > 1 then
5    ∃m ∈ F: ∀n ∈ F, n ≠ m:
     m.time > n.time ⇒ execute
     compensating activity c(π2(m))
6   else
7    ∃m ∈ F ⇒ execute compensating
     activity c(π2(m))
8   F ← F \ {m}
9   ∀v ∈ o(c(π2(m))): V_result ← V_result ∪ {(v,
    c, t)}, c is the new value of variable
    v, t is the timestamp of the
    assignment
10 od
11 return V_result
```

A similar algorithm for the creation of the reverse order graph is also proposed in [17]. But the intention of our algorithm is to deliver the changed variable values as result of the compensation operation.

Definition 17 (Re-execute Operation). The re-execution is a function that repeats logic of a process model for a given process instance with a given activity as starting point. The data indicated by the given time step is taken as input for the re-execution. The operation uses the compensate operation for already completed work in the iteration body.

$\mathfrak{r}$: A $\times$ $\mathcal{P}_{all}$ $\times$ $\mathbb{N} \mapsto \mathcal{P}_{all}$

$a \in$ A, p_in_g, $p\_out_g \in \mathcal{P}_{all}$ and the pre-condition are similar to the iterate operation. The difference is the calculation of $\mathcal{V}_{out}$: $\mathfrak{r}(a, p\_in_g, t) = p\_out_g$: $\Leftrightarrow$

1. $\mathcal{V}_{out} = \mathcal{V}_{in} \cup$ compensateIterationBody(*a*, *p*)

The variable values might be modified as a result of the compensation of completed work in the iteration body (1.). Note that the start activity for the re-execution is scheduled after the compensation is done.

C. Data Handling

For the repetition of workflow parts the handling of data is of utmost importance. Where to store data that the former iteration has produced? What data should be taken as input for the next iteration? A mechanism is needed to store different values for same variables and to load variable values for iterations. The compensation of completed work as is done in the re-execution operation is not sufficient for resetting variables because compensation does not always set the variables to their former states. This strongly depends on the compensation logic and invoked services.

The desired functionality can be realized by saving the complete content history of variables including the assignment timestamps. Many workflow systems already provide this as part of their audit trail [2, 18]. If a variable is changed a new tuple is inserted in $\mathcal{V}$ and the former tuples remain, e.g., at time step 9 variable *number* was increased by 1: $\mathcal{V}$ = {(number, 50, 1), (number, 51, 9)}. That way no data is lost due to repetition of workflows parts and former variable values can be accessed as input for the rerun. It must be possible for the users to choose the input for the next iteration. That is why we foresee the specification of a time step in the iteration and re-execution operation (see Definition 15 and 17). This time step indicates which variable values are taken as input for the respective operation, namely those that were valid at the given point in time. The visible variables have to be re-initialized accordingly. In Figure 3, the sample workflow of Figure 1 was iterated from activity *a* three times leading to a chain of executions of activities *a* and *b*. The current value of variable *number* was taken for each rerun. Imagine the user wants to iterate again from activity *a*. Different time steps chosen by the user as input for the operation would influence the initialization of variables for the iteration as follows. At the (latest) time t = 8 the value of *number* is 102; at t = 6 the value is the one obtained after the second execution of *a* (101); and at t = 1 *number* has its initial value (99).

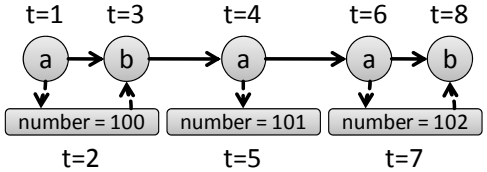

Figure 3. Data handling during workflow repetition

D. Implications on the Execution Correctness

In practice, workflows consist of different activity types, e.g., for sending/receiving messages, loops. The enforced repetition of workflow logic has to account for different activity types, especially those that interact with external

entities such as clients, humans, services/programs. The main problem is that the repetitions are not reflected in the workflow logic and hence the aforementioned entities do not know a priori the exact behavior of the workflow.

If a message receiving activity is repeated, the message sending client has to re-send the message or send an adapted message. The problem is that the client needs to be informed about the repetition. A simple solution is that clients provide a special operation that can be used by the workflow engine to propagate the iteration. Over this operation the engine could send the message back to the client enriched with further context (e.g., the address of the engine, correlation information, workflow instance id). The client then decides whether to re-send the message or to send an adapted one.

The repetition of message sending activities is straight forward for idempotent services. Non-idempotent services should be compensated prior to a repeated invocation, as is done in the re-execution operation. If the iteration operation repeats the execution of non-idempotent services, then the user is responsible for the effect of the operation.

Iterations within modeled loops can have an unforeseeable impact on the behavior of workflows. The context of workflows might be changed in a way that leads to infinite loops (e.g., because the repetition changes variable values so that a while condition can never evaluate to false). Usually, a workflow system provides operations to change variable values. This functionality can be used to resolve infinite loops.

E. User Interaction With the Workflow System

A workflow system that implements our approach must provide a monitoring tool that allows users to continuously follow the execution state of process instances. If the user detects a faulty or unintended situation, he can suspend the workflow and manually trigger an iteration/re-execution. The system requests him for the time step used to retrieve the variable values for the loop. Then, the process instance state is changed as described in Section IV and the user can resume workflow execution.

V. CONCLUSION AND OUTLOOK

In this paper, we have formally described two operations to enforce the repetition of workflow logic during workflow runtime: the *iterate* operation reruns activities starting from a manually selected activity; the *re-execute* operation undoes completed work in the iteration body before rerunning activities. The distinctive features of the approach are that the repetition does not have to be modeled or configured previously and that arbitrary activities can be used as starting point for the rerun. We have shown how problems with the data handling and communication with external parties can be solved. The approach is described based on an abstract meta-model and thus can be applied to existing or future workflow engines and languages. Currently, we are working on an implementation for the BPEL engine Apache ODE.

The enforced repetition of workflow logic is a step towards our goal to enable an explorative workflow development, especially in the field of scientific workflows.

ACKNOWLEDGMENT

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

REFERENCES

[1] W. van der Aalst, T. Basten, H. Verbeek, P. Verkoulen, and M. Voorhoeve, “Adaptive workflow: on the interplay between flexibility and support,” Proc. of the 1st Conf. on Enterprise Information Systems, pp. 353-360, 1999.

[2] F. Leymann and D. Roller, “Production Workflow – Concepts and Techniques,” Prentice Hall, 2000.

[3] I. Wassink, M. Ooms, and P. van der Vet, “Designing workflows on the fly using e-BioFlow,“ ICSOC, 2009.

[4] R. Barga and D. Gannon, “Scientific vs. business workflows,” in: Taylor et al., “Workflows for e-Science,” Springer, pp. 9-18, 2007.

[5] H. Eberle, O. Kopp, F. Leymann, and T. Unger, “Retry scopes to enable robust workflow execution in pervasive environments,” Proc. of the 2nd MONA+ Workshop, 2009.

[6] Oracle BPEL Process Manager, http://www.oracle.com/us/products/middleware/application-server/bpel-home-066588.html

[7] E. Deelman, G. Mehta, G. Singh, M.-H. Su, and K. Vahi, “Pegasus: Mapping large-scale workflows to distributed ressources,” In: Taylor et al., “Workflows for e-science,” Springer, pp. 376-394, 2007.

[8] G. Vossen and M. Weske, “The WASA approach to workflow management for scientific applications,” Workflow Management Systems and Interoperability, NATO ASI Series F: Computer and System Sciences, Vol. 164, Springer-Verlag, pp. 145-164, 1998

[9] F. Leymann, “Supporting business transactions via partial backward recovery in workflow management systems,” Proc. of the BTW, Springer, 1995.

[10] Object Management Group (OMG), “Business Process Modeling Notation (BPMN) Version 1.2,” OMG Specification, 2009.

[11] M. Reichert and P. Dadam, “$ADEPT_{flex}$ – Supporting dynamic changes of workflows without losing control,“ Intelligent Information Systems, vol. 10, pp. 93-129, 1998.

[12] D. Chiu, Q. Li, and K. Karlapalem, “A meta modeling approach to workflow management systems supporting exception handling,” Information Systems, vol. 24, pp. 159-184, 1999.

[13] Apache ODE, http://ode.apache.org/activity-failure-and-recovery.html

[14] P. Greenfield, A. Fekete, J. Jang, D. Kuo, “Compensation is not enough,” Proc. of the 7th EDOC, 2003.

[15] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, “Taverna: a tool for building and running workflows of services,” Nucleic Acids Research, vol. 34, Web Server issue, pp. 729-732, 2006.

[16] I. Altintas, O. Barney, E. Jaeger-Frank, “Provenance Collection Support in the Kepler Scientific Workflow System,“ Provenance and Annotation of Data, IPAW, LNCS, Vol. 4145, pp. 118–132, Springer, 2006.

[17] R. Khalaf, “Supporting business process fragmentation while maintaining operational semantics: a BPEL perspective,” Doctoral Thesis, ISBN: 978-3-86624-344-6, 2008.

[18] Workflow Management Coalition, “Audit Data Specification, Version 1.1,” WfMC Specification, 1998.

 ISBN: 978-1-61208-160-1

Towards "Executable Reality": Business Intelligence on Top of Linked Data

Vagan Terziyan
Department of Mathematical Information Technology,
University of Jyvaskyla
P.O. Box 35 (Agora), 40014, Jyvaskyla, Finland
e-mail: vagan@jyu.fi

Olena Kaykova
Industrial Ontologies Group, Agora Center,
University of Jyvaskyla
P.O. Box 35 (Agora), 40014, Jyvaskyla, Finland
e-mail: olena@cc.jyu.fi

Abstract—**Tight competition within mobile technology domain resulted to quite advanced applications and services for the users. Among them there are mobile (augmented and mixed) realities as an effort to enhance real-world observation by bridging it with virtual worlds of relevant data and services. At the same time, due to emerging Semantic Technology, the Web content moves rapidly towards Linked Data. The layer of machine-processable semantics allows automated processing of the content by Web-based Business Intelligence (BI) applications and services. Real-time analytics related to various real-life objects provided to the users of Mixed Reality by online BI services would be a nice enhancement of the technology. In this paper we propose "Executable Reality" as an enhancement of the "Mixed Reality" concept within two dimensions (utilization of Linked Data and BI on top of it). We present "Executable Knowledge" as a tool to enable Linked Reality and "Executable Focus" to control it by a user. Executable knowledge in addition to subject-predicate-object semantic triplet model (in ontological terms) contains also subject-predicate-query triplets. Actual value for the properties based on a new triplet will be computed "on-the-fly" (based on user request navigated by executable focus) by some online BI service or other computational capability provider at a right time and place according to the dynamic user context.**

Keywords- Business Intelligence; Linked Data; Mixed Reality; Executable Reality; Executable Knowledge

I. INTRODUCTION

Business intelligence (BI) can be considered as a set of methods, techniques and tools utilized on top of business data to compute (acquire, discover) additional (implicit) analytics out of it and to present it in a form suitable for decision-making, diagnostics and predictions related to business. Taking into account that "business data" is becoming highly heterogeneous, globally distributed (not only in the Internet space but also in time), huge and complex, extremely context sensitive and sometimes subjective, the ways the BI is utilized have to be qualitatively changed. Semantic (Web) Technology [1, 2, 3] is known to be a suitable approach to enable more automation within BI-related data processing. The vision of BI 2.0 [4] includes also issues related to SOA, mobile access, context handling, social media, etc. All these issues will also definitely benefit from adding semantics [5, 6]. It is however a known fact that there is not much semantically annotated data available for BI. We have to live with data sets created independently, according to different schemas or even data model types. The realistic role of Semantic Technology for such data would be linking related "pieces" of it with some semantic connections and by doing this transforming the original data into Linked Data.

There are no doubts that such semantically interlinked "islands" of data have a lot of hidden (implicit) and potentially interesting information that none of separate data sets has alone. Now the challenge would be how to utilize BI on top Linked Data to be able to get all the benefits from semantic enhancement of the data.

Another trend is related to fast development of technology, tools and devices for better delivery and visualization of information to a user. Among those there are technologies like Augmented Reality [11], Mixed Reality [14] and mobile versions of both [12, 13, 15]. These technologies are based on automated interlinking of various Web-based digital data collections with the real-time data from sensors about physical world and presenting it in a useful form for a user. An interesting topic would be considering these technologies in the context of business data or even BI-provided analytics. This may inspire more professional use of Augmented and Mixed Reality in addition to public use of it.

In this paper we propose "Executable Reality" as an enhancement of the "Mixed Reality" concept within two dimensions (utilization of Linked Data and BI on top of it). We present "Executable Knowledge" as a tool to enable Linked Reality and "Executable Focus" to control it by a user. Executable knowledge in addition to subject-predicate-object semantic triplet model (in ontological terms) contains also subject-predicate-query triplets. Actual value for the properties based on a new triplet will be computed "on-the-fly" (based on user request navigated by executable focus) by some online BI service or other computational capability provider in a right time and place and according to dynamic user context.

The rest of the paper is organized as follows: in Section II we discuss Linked Data issues and its enhancement by context-sensitive similarity links; in Section III we present (Mobile) Augmented and Mixed Reality technology and challenges; In Section IV we show how these technologies can be further developed towards "Executable Reality" on top of enhanced Linked Data and BI services (there we also present concept of "Executable Knowledge"); we discuss Related Work in Section V; and we conclude in Section VI.

 ISBN: 978-1-61208-160-1

II. Linked Data and Semantic similarity

Linked Data is a concept closely related to the Semantic Web yet providing some specific facet into it. According to Tim Berners-Lee "The Semantic Web is not just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data" (http://www.w3.org/DesignIssues/LinkedData.html). The so called "5 stars" advice from Tim Berners-Lee to enable Linked Data includes: making data available on the web (whatever non-proprietary format) as machine-readable structured data, utilizing open standards from W3C (RDF and SPARQL) to identify things and finally linking the data to other people's data to provide context.

According to Kingsley Idehen (OpenLink Software CEO), due to development of Semantic Technology, *meshing* (or natural data linking) will replace *mashing* (brute-force data linking) and therefore mesh-ups can be considered as a next step comparably to the mash-ups in the sense to merge and integrate different data sources and processing devices to provide new information services.

Linked Data can be considered as an outcome of the technology, which semantically interconnects heterogeneous data "islands" (e.g., as shown in Figure 1). Even if the original sources of data are highly heterogeneous (not just only different schema of data within the same data model type but also different data model types), still it is possible to build some "bridges" between entities from these data sources utilizing semantic technology. The traditional Semantic Web approach would be: (a) creating semantic model of the domain (ontology), (b) replacing original data from each source with full semantic (RDF) representation of its resources in terms of the ontology. Of course such approach enables seamless integration of the original data and simplifies the usage of it. However with distributed and dynamic sources of data, which are managed and constantly updated independently, it would be difficult to provide such "semantic synchronization" (updating metadata and mapping it to the ontology) in real time. Therefore Linked Data would be less ambitious and more practical approach: data sources are managed independently as they used to be; semantic connections between appropriate resources from different sources will be either automatically discovered or manually created whenever appropriate. Usage experience and usability performance for each separate data source will be preserved. The usability of such "virtually integrated" data storages will increase with the increase of the amount of the semantic "bridges".

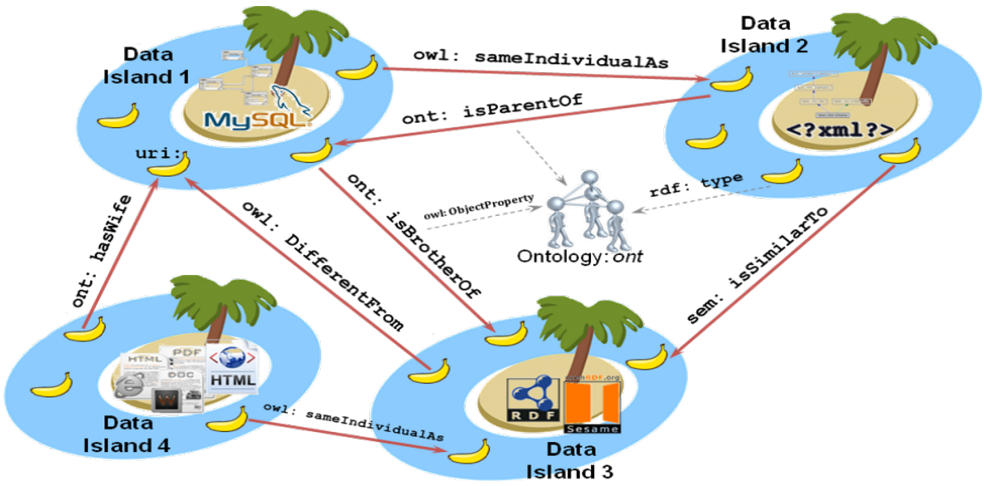

Figure 1. Linked Data: "bridges" between heterogeneous "islands" of data

According to [7] there are three important types of RDF links within Linked Data: (a) "relationship links" that point at related things in other data sources (like "object properties" in terms of OWL: *owl: ObjectProperty*); (b) "identity links" that point at URI aliases used by other data sources to identify the same real-world object or abstract concept (e.g., *owl: sameIndividualAs*, *owl: equivalentClass*); (c) vocabulary links that point from data to the definitions of the vocabulary terms that are used to represent the data (like "datatype properties" in terms of OWL: *owl: DatatypeProperty*).

We think it would be reasonable to extend traditional explicit semantic links within Linked Data with the implicit ones, e.g., those, which could be automatically derived by various reasoners. Among those special attention should be paid to the "semantic similarity" links. Usually, when one queries a data, she looks for the resource(s), which are "the same" as the one specified in the query. However often there are none of such found. In many cases there is a sense to find "similar" to a target one resources. Similarity search was always a big issue within many disciplines and it is especially important for the Linked Data. The reason is related to the fact that actually same resources in different "islands" of data may have different URIs and often quite extensive work should be done to recognize same resources. Usually first we see that some resources look similar and

 ISBN: 978-1-61208-160-1

therefore in practice could be the same ones and then we perform some check on the identity of the resources. Another use of similarity measure is when one is looking for a capability providing resource (e.g., a service), cannot find exactly the one she wants but still will be satisfied by finding a resource with similar functionality. Therefore explicit similarity links between data entities would be reasonable to have as a result of appropriate similarity search procedures. The two major challenges here are: (a) a resource within one data "island" may have very different model of description when compared to some resource within another data "island" (e.g., a human documented in a relational database will not be easily compared with a human from some XML storage or from some html document); (b) some resources being very different in one particular context could be considered as similar ones in some other context.

We consider three types (sub-properties in terms of OWL) of semantic similarity based on common ternary object property relation named *isSimilarGivenContext* and they are: (1) *isSimilarGivenQuery*; (2) *isSimilarGivenRole*; and (3) *isSimilarGivenGoal*. The first type of similarity assumes that two resources can be considered as similar ones (in the context of some semantic query, e.g., SPARQL query) if this query, being applied over the locations of these two resources, returns both of them as a result. For example, resources "Mikhail S. Gorbachev" and "George W. Bush" will be considered as similar ones, given query "Former president, male with at least one daughter". The second type of similarity assumes that two resources will be used as similar ones if they both can fill some slot in a business process with the specified role. For example, some instance of class "Lamp" can be considered similar to some instance of class "Candle", given role "Lightening". The third type of similarity applies to the resources which can be replaced with each other as input parameters needed to perform some function (action) or utilize some external capability (service) without affecting expected outcome. For example, a "Bottle of vodka" would be a similar resource to e.g., "Pack of beer" as an "input" ("natural payment") to a ferry service somewhere in Siberia countryside, given goal "To cross the river". More information about our approach for defining context in various practical applications and semantic similarity search within context can be found in [8, 9, 10].

The major challenge is how to provide support for automated utilization of Linked Data, which in fact remains heterogeneous, and how to get added value of additional semantic connections between data components. Anyway we claim that providing similarity links, in addition to traditional types of RDF links described in [7], can be very helpful for practical utilization of Linked Data and we will try to show this in the following sections of the paper.

III. Augmented and Mixed Reality

Augmented Reality (AR) [11] is a technology aimed to enhance the traditional perception of a reality (real-world environment), which elements are augmented by computer-generated sensory input (e.g., data, sound or graphics). AR enriches real world for the user rather than replaces it. By contrast, *virtual reality* replaces the real-world with a simulated one. Emerging development of mobile computing has naturally resulted to growing interest towards Mobile AR [12] and also to Ubiquitous Mobile AR [13] for successfully bridging the physical world and the digital domain for mobile users. The AR concept has been further developed to Mixed Reality [14], which means merging of real and virtual worlds to produce new environments and visualizations where physical and digital objects co-exist and interact in real time. In June 2009, Nokia Research Center announced the vision [15] of Mobile Mixed Reality, according to which a phone becomes a "magic lens" (smart and context-aware), which lets users look through the mobile display at a world that has been supplemented with information about the objects that it sees. The users simply point their phone's camera, and look "through" the display. Objects of interest visible in the current view will be gathered from existing Point-Of-Interest databases or created by the user and will be highlighted. They can be associated with physical objects or featureless spaces like squares and parks. Once selected, objects provide access to additional information from the Internet and hyperlinks to other related objects, data, applications and services. Context-awareness is guaranteed by various rich sensors that are being incorporated into new phones (GPS location, wireless sensitivity, compass direction, accelerometer movement, sound and image recognition, etc.). Therefore the new technology is going to actively utilize acquired dynamic context to better filter and select relevant information about surrounding real-world objects for a user.

In the following section we further develop the concept of (mobile) mixed reality within two dimensions: the first one is related to Linked Data utilization and the second one will be related with the utilization of Business Intelligence through "Executable Knowledge".

IV. Towards "Executable" Reality

The concept of "Executable Reality" and associated technology, which we are offering, should be considered as an extension of the (Mobile) Mixed Reality concept and the technology. If the traditional technology assumes that the explicitly available relevant data about some real-world object will be taken from some database and delivered to the user on demand (based on her attention focus pointed to this object), the Executable Reality in addition is able to provide online BI computation on similar demand (we call it "Executable Focus") and present to the user results of computed analytics adapted for the current context. Two use-case scenarios for the Executable Reality are shown in Figure 2 (a, b). In the first one, the user (maintenance engineer of the power network company) is putting the executable focus (smart phone camera) into direction of the power line and by doing this makes implicit request for the last 24 hours performance statistics of this power line. The query will go further to server; appropriate BI service will be selected and automatically invoked; resulting page with numbers, graphics (and sounds if appropriate) will be generated and delivered back to the terminal and shown in appropriate window of the screen as shown in Figure 2 (a).

 ISBN: 978-1-61208-160-1

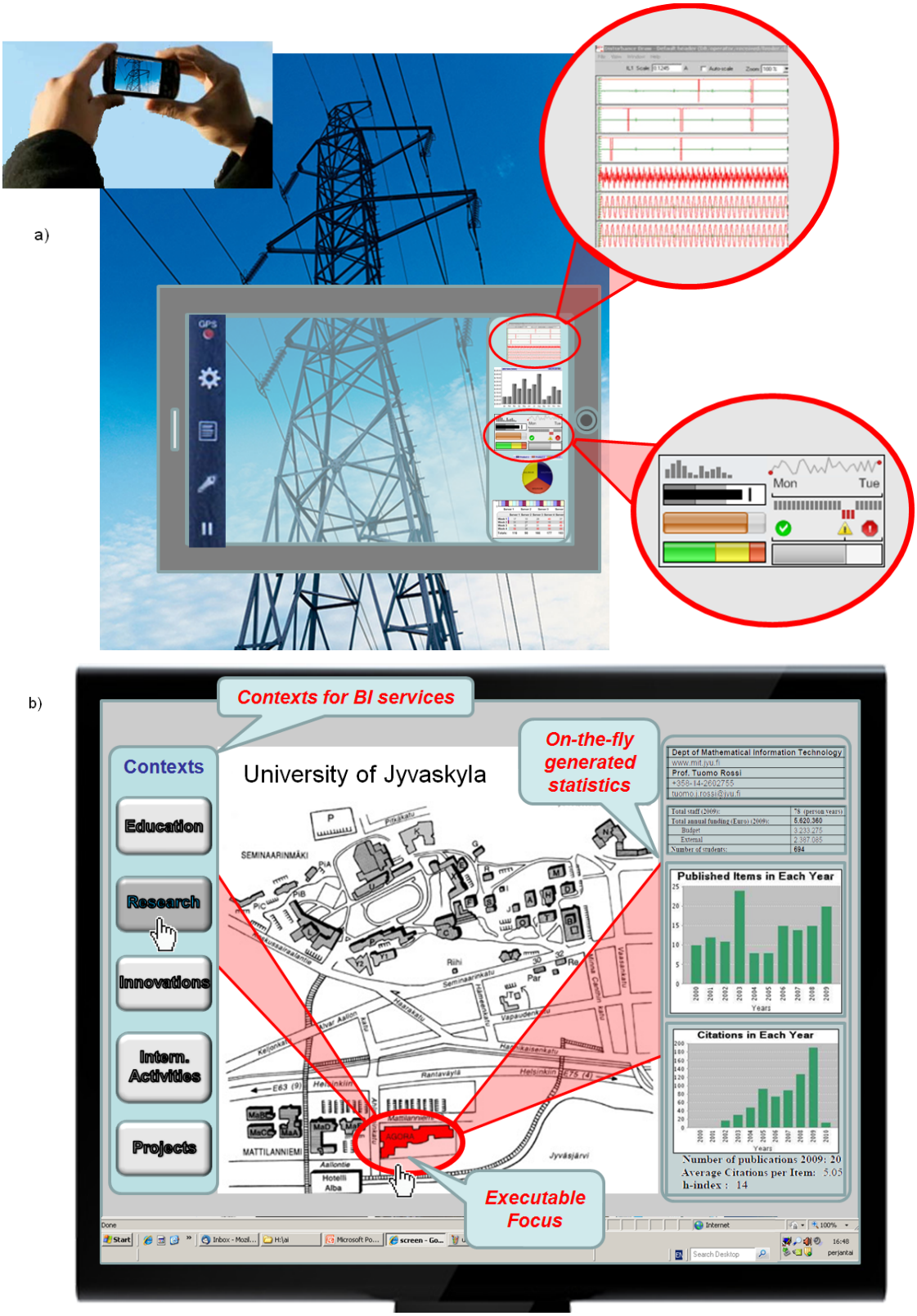

Figure 2. Executable Reality use case examples: (a) on-the-fly computed statistics about power line performance is delivered to mobile terminal of the maintenance engineer on implicit demand; (b) research performance statistics is delivered to the user based on chosen unit (click on the building where the university department is located and selecting context “research” for filtering appropriate data from the unit needed for research performance calculations)

 ISBN: 978-1-61208-160-1

Another scenario in Figure 2 (b) shows that a user observes the campus map of some university and selects the building where a particular department is located. The user also selects the context in which she wants to get performance statistics of the unit, e.g., "research". Chosen object and the context together form the executable focus, which will automatically generate the query for the required computation. Then the process goes in a similar way as with previous scenario and the user will get "fresh" statistics concerning research performance of the department chosen.

To enable Executable Reality we have to find effective way to utilize Linked Data, which is natural source for online BI computations, and also to enable BI functionality as semantically annotated Web services. We propose to organize Linked Data in form of "Executable Knowledge".

Executable Knowledge can be considered as distributed (or organized as a cloud) set of heterogeneous data storages and computational services (e.g., BI) interconnected with semantic (RDF) links. The major feature here is that, in addition to the traditional ("subject-predicate-object" or "resource-property-value") triplet-based semantics of an RDF link (either "datatype" property, where "value" is a literal; or "object" property, where "value" is another resource), the new model will have new property type named "executable property" with the structure: "subject-predicate-query". It is supposed that reasoners, engines, etc. working with such knowledge will execute query within target triplet and treat obtained result as a value for the property. Two immediate advantages of this extension are: (a) triplet will always implicitly keep knowledge about most recent value for the property because query to some data storage or to some BI function will be executed only on demand when needed and the latest information will be delivered; (b) query may be written according to different standards, data representation types, models and schemas so that heterogeneity of original sources of data and capabilities will not be a problem. Therefore distributed data collections can be maintained independently (autonomously) and "queried" in real time by executable RDF links.

Consider simple example in Figure 3. Here the executable statement in RDF (N3 syntax) actually means: "If you want to know with whom John is currently in love, execute the query Q1". The query Q1 (prefix "exe:" points to Executable Knowledge ontology of various query types and indicates that the RDF statement is executable) in this case is semantically described as SPARQL query to the RDF data storage and it means: "Select the girl from the current database of staff, who is colleague of John, has red hair and is 25 years old". When the SPARQL query engine executes the query and finds that "Mary" fits it, the executable RDF statement is transformed into the traditional one (reference to the query "exe:Q1" is replaced with actual value "Mary"). Notice that, if the same knowledge will be explored after 1 year, then the same executable statement will be transformed into: "John is in love with Anna", because staff data (separate source) will be autonomously updated (Anna becomes 25 years old) and RDF connections (semantic layer of Linked Data) on top will be updated when executed.

Consider similar example in Figure 4 and notice that here we have an SQL query to some relational database as implicit value of executable RDF statement. The query Q2 means request for computing average journal papers' publication performance of young (< 30) PhD students. The original executable RDF statement means: "If you want to know average performance of young doctoral students in AI Department, execute query Q2". When the query returns computed value the executable RDF statement is transformed into the traditional one (reference to query "exe:Q2" is replaced with actual value "7", which means that "executable" RDF property is replaced with "datatype" one).

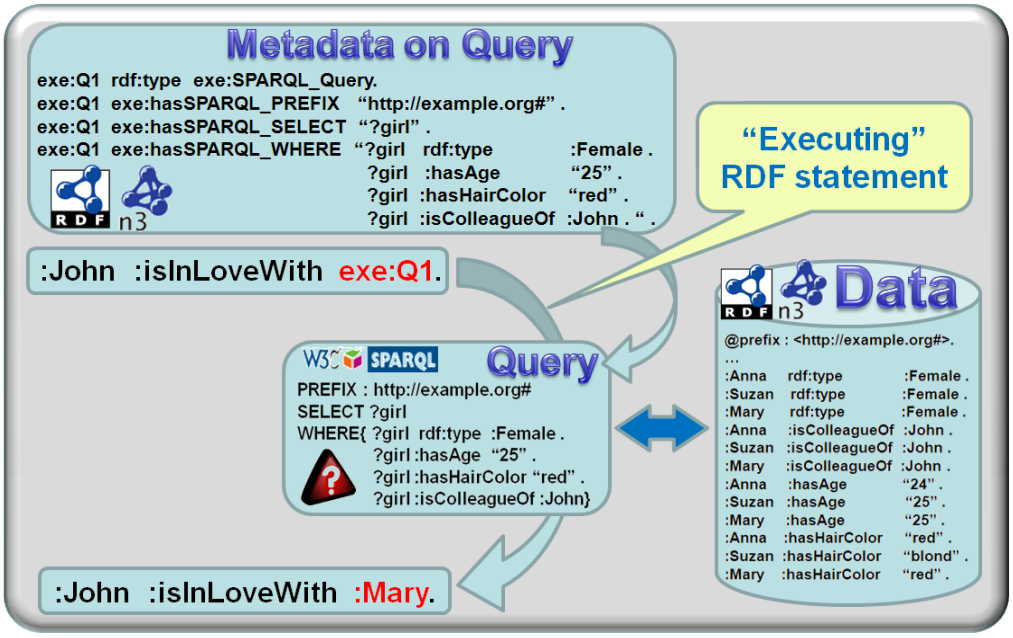

Figure 3. Example of processing executable RDF statement, which contains implicit value as SPARQL query to RDF storage

 ISBN: 978-1-61208-160-1

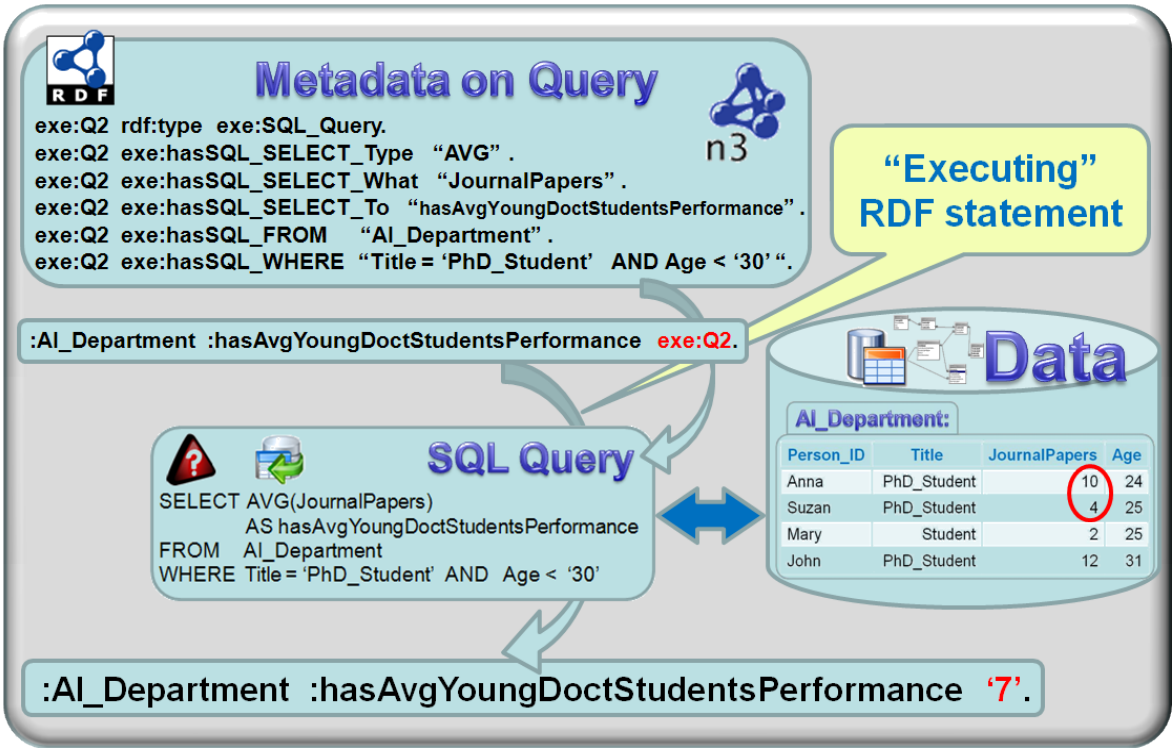

Figure 4. Example of processing executable RDF statement, which contains implicit value as SQL query to a relational database

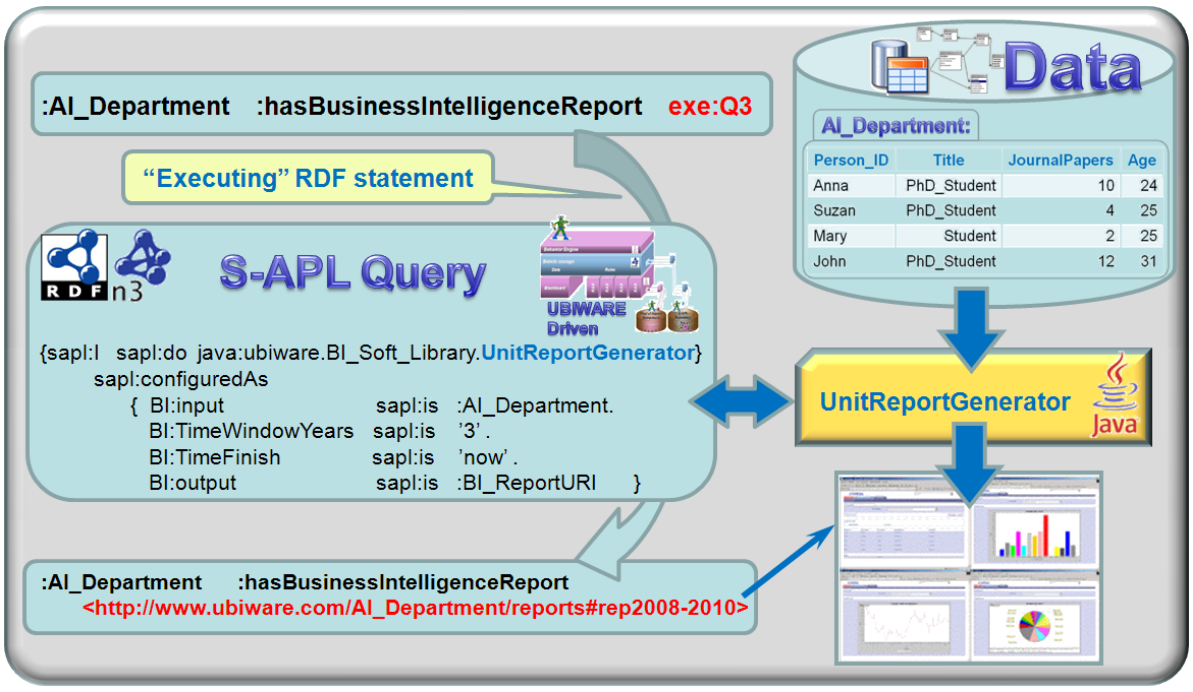

Figure 5. Example of processing executable RDF statement, which contains implicit value as S-APL query to BI software as a service

Consider example in Figure 5. Here we have executable RDF statement that can be interpreted like: "If you want to get basic BI-statistics report for the AI Department for last 3 years, execute query Q3". Behind this query there is a Java software module "UnitReportGenerator" provided as-a-service from online software library. The query itself is written in S-APL (Semantic Agent Programming Language [16]) used for UBIWARE-based applications [17]. S-APL is

 ISBN: 978-1-61208-160-1

RDF-based language for multi agent systems, in which both data and actions are described semantically. UBIWARE [18] ("Smart Semantic Middleware for Ubiquitous Computing") has been developed by Industrial Ontologies Group (http://www.cs.jyu.fi/ai/OntoGroup). It as a software technology and a tool to support design and installation, autonomic operation and interoperability among complex, heterogeneous, dynamic and self-configurable distributed systems, and to provide coordination, collaboration, interoperability, data and process integration service. UBIWARE platform is used actually to deal with "Executable Knowledge" and its utilization for "Executable Reality" services. In the example, when the BI software is executed, it generates the html page where all analytics is visually presented with different BI widgets. The URI for this page will replace the implicit "exe:Q3" value from the original RDF statement and creates traditional RDF statement with object property connecting two resources (AI Department URI and BI statistics Report URI).

There is also a possibility to compute semantic similarity between resources from different data storages and automatically create appropriate RDF connections for similar (same) instances. As it was shown in Section II, some instances can be considered as similar ones in one context and can be considered otherwise in other context. Therefore, similarly to the examples above, the "Executable Knowledge" supports also RDF statements with implicit similarity search queries, in which needed query parameters are automatically taken from the current context. Change of context can be considered as implicit query (if appropriate setup is made) to re-compute similarity links, which makes RDF graph on top of Linked Data very dynamic. Our approach for context-sensitive semantic similarity computing and its implementation is discussed in [10].

Mixed Reality is just one possible way to utilize Executable Knowledge concept. There should be definitely other application areas for it. Generally many industrial applications, which require dynamic self-configurable solutions, applications and architectures, will benefit from the flexible Executable Knowledge, as our experience with UBIWARE industrial cases demonstrates [18].

V. Related Work

Concepts of "virtual", "augmented", "mixed", etc., realities discussed in Section 1 are being actively developed into various services for public. There are many other relevant concepts and activities, which have many common features with the above, having however some specifics. One such abstraction is so-called "Mirror World" [19], which is a representation of the real world in digital form mapped in a geographically accurate way. Mirror worlds are can be seen as an autonomous manifestation of digitalized reality including virtual elements. Other relevant concept is "Metaverse" (http://metaverseroadmap.org), which is the convergence of virtually-enhanced physical reality and physically persistent virtual space being a fusion of both. The "Second Life" (http://secondlife.com/) is a 3D virtual world enhanced by social networks and communication capabilities. "Lifelogging" [20] is continuous capturing from a human and sharing through the Web various data, events and activities collected by various devices, sensors, cameras, etc. Other slightly different concept is "Lifeblog" [http://europe.nokia.com/support/product-support/nokia-photos], which is also known as popular service for collecting and putting into a timeline (mobile) user activities and creating data in the form of complex multimedia diary.

Our intension was to find out reasonable services out of these concepts suitable not just for public use but mostly for professionals. For that we explored the possibility to utilize Business Intelligence as an additional capability. Preliminary information on the interesting relevant effort named "Augmented Business Intelligence" has appeared in the Web [21] just a couple of months ago. Augmented BI is considered in [21] as a process of using a mobile device to scan an image or a barcode and overlaying metrics and or charts onto the image. This supposes to facilitate the process of a store manager moving around a retail store and would like more information about that products sales performance. See Figure 6, which demonstrates possible use case for the Augmented BI. There are some evident similarities with our use-cases shown in Figure 2, however our implementation benefits from the Linked Data utilization and allows context-sensitive view to the BI-enhanced reality.

Figure 6. Demonstration of possible Augmented Business Intelligence utilization scenario [21].

Our solution related to BI-enhanced mixed reality (or Executable Reality) is based on Executable Knowledge concept. The Executable Knowledge inherits some features from a Dynamic Knowledge (see, e.g., [22]), which is actually dynamically changing knowledge and according to (www.imaginatik.com) providing on-demand, in-context, timely, and relevant information. Issues related to such knowledge include power and expressive tools and languages (as, e.g., LUPS [23]) for representing such knowledge and proper handling of conflicting updates as addressed in [22]. Given an initial knowledge base (as a logic program) LUPS provides a way for sequentially updating it.

Since executable knowledge is definitely a kind of dynamic knowledge, other issue would be whether it is declarative or procedural knowledge. A procedural knowledge (or knowledge on how to do something) is

 ISBN: 978-1-61208-160-1

known to be a knowledge focused on obtaining a result and exercised in the accomplishment of a task, unlike declarative knowledge (propositional knowledge or knowledge about something) [24]. Procedural knowledge is usually represented as finite-state machine, computer program or a plan. It is often a tacit knowledge, which means that it is difficult to verbalize it and transfer to another person or an agent. The opposite of tacit knowledge is explicit knowledge.

The concept of executable knowledge can be actually considered as a kind of hybrid of declarative and procedural knowledge. As it can be seen from the examples in Section 4, by "executing" knowledge one actually transforms tacit (procedural) knowledge into explicit (declarative one). Therefore an executable knowledge contains explicit procedural (meta-) knowledge on *how to acquire* (or compute) declarative knowledge. Such capability means that the executable knowledge is naturally self-configurable knowledge (or more generally – self-managed knowledge). We use S-APL (Semantic Agent Programming Language [16]) for its representation, which is based on RDF (N3) syntax and which is equally suitable to manage declarative and procedural knowledge.

Our implementation of the executable knowledge on top of UBIWARE [17,18] agent-driven platform allows UBIWARE agents autonomously "execute" knowledge by following explicit procedural instructions for BI services execution and therefore updating (or making explicit) appropriate declarative beliefs.

VI. Conclusions

In this paper, we presented one way on how Linked Data (from heterogeneous sources) can be automatically processed by various BI services; and also how the results of BI processing can be shown through the (Mobile) Mixed Reality technology. Data heterogeneity problem is handled by the "Executable Knowledge" approach, according to which semantic (RDF) links include explicit queries to data or to (BI) services and other capabilities based on various possible data models and the context.

References

[1] Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web. Scientific American, 284(5), 2001, pp. 34–43.

[2] Sheth, A. and Ramakrishnan, C., Semantic (Web) Technology in Action: Ontology Driven Information Systems for Search, Integration and Analysis, IEEE Data Eng. Bulletin, 26(4), 2003, pp. 40-48.

[3] Hitzler, P., Krötzsch, M. and Rudolph, S., Foundations of Semantic Web Technologies, Chapman&Hall/CRC, 2009, 455 pp.

[4] Nelson, G., Business Intelligence 2.0: Are we there yet? In: Proceedings of the SAS Global Forum 2010, Seattle, USA, 11-14 April, 2011, paper 040-2010.

[5] Domingue, J., Fensel, D., and González-Cabero, R., SOA4All, Enabling the SOA Revolution on a World Wide Scale, In: Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC 2008), August 4-7, 2008, Santa Clara, California, USA, IEEE CS Press, 2008, pp. 530-537.

[6] Sell, D., Cabral, L., Motta, E., Domingue, J. and Pacheco, R., Adding Semantics to Business Intelligence, In: Proceedings of 16th International Workshop on Database and Expert Systems Applications, Copenhagen, 26 August, 2005, pp. 543 – 547.

[7] Heath, T., and Bizer, C., Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011, 136 pp.

[8] Khriyenko, O., and Terziyan, V., A Framework for Context-Sensitive Metadata Description, International Journal of Metadata, Semantics and Ontologies, 1(2), 2006, Inderscience Publishers, pp. 154-164.

[9] Terziyan, V., Predictive and Contextual Feature Separation for Bayesian Metanetworks, In: B. Apolloni et al. (Eds.), Proceedings of KES-2007 / WIRN-2007, Vietri sul Mare, Italy, September 12-14, Vol. III, Springer, LNAI 4694, 2007, pp. 634–644.

[10] Khriyenko, O., Terziyan, V., Similarity/Closeness-Based Resource Browser, In: J.J. Zhang (Ed.), Proceedings of the Ninth IASTED International Conference on Visualization, Imaging and Image Processing (VIIP-2009), July 13-15, 2009, Cambridge, UK, ACTA Press, pp. 184-191.

[11] Azuma, R., A Survey of Augmented Reality, Presence: Teleoperators and Virtual Environments 6 (4), 1997, MIT Press, pp. 355-385.

[12] Hollerer, T., Feiner, S., Terauchi, T., Rashid, G., Hallaway, D., Exploring MARS: Developing Indoor and Outdoor User Interfaces to a Mobile Augmented Reality System, Computers & Graphics, 23, 1999, Elsevier, pp. 779-785.

[13] Henrysson, A., Ollila, M., UMAR: Ubiquitous Mobile Augmented Reality, In: Proceedings of the Third International Conference on Mobile and Ubiquitous Multimedia (MUM-2004), College Park, Maryland, USA, 2004, pp. 41-45.

[14] Ohta, Y., Tamura, H. (Eds.), Mixed Reality: Merging Real and Virtual Worlds, 1999, Springer, 418 pp.

[15] Mobile Mixed Reality: The Vision, Nokia Technology Insights Series, Nokia Research Center, June 2009, 4 pp., Available online in: http://research.nokia.com/files/insight/NTI_MARA_-_June_2009.pdf.

[16] Katasonov, A. and Terziyan, V., SmartResource Platform and Semantic Agent Programming Language (S-APL), In: P. Petta et al. (Eds.), Proceedings of the 5-th German Conference on Multi-Agent System Technologies (MATES'07), 24-26 September, 2007, Leipzig, Germany, Springer, LNAI 4687 pp. 25-36.

[17] Katasonov, A., Terziyan, V., Implementing Agent-Based Middleware for the Semantic Web, In: Proceedings of the Second IEEE International Conference on Semantic Computing (ICSC-2008) / International Workshop on Middleware for the Semantic Web, August 4-7, 2008, Santa Clara, USA, IEEE CS Press, pp. 504-511.

[18] UBIWARE: http://www.cs.jyu.fi/ai/OntoGroup/UBIWARE.htm , Project Web Site, Industrial Ontologies Group, 2008-2011.

[19] Gelernter, D., Mirror Worlds: or the Day Software Puts the Universe in a Shoebox...How It Will Happen and What It Will Mean, 1st ed., Oxford University Press, 1992.

[20] O'Hara, M., Tuffield, M., Shadbolt, N., Lifelogging: Privacy and Empowerment with Memories for Life, In: Identity in the Information Society, Vol. 1, No.1, Springer, pp. 155-172.

[21] Husting, P., Augmented Business Intelligence in Retail, In: Microsoft BI Collaboration and Community Blog, 3 March 2011, Available online in: http://blog.extendedresults.com/2011/03/03/augmented-business-intelligence-in-retail/ (accessed 15 June 2011).

[22] Alferes, J., Pereira, L., Przymusinska, H., Przymusinski, T., Quaresma, P., Dynamic Knowledge Representation and its Applications, In: Proceedings of the 9th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA '00), Varna, Bulgaria, September 20-23, 2000, LNCS, Vol. 1904, Springer, pp. 1-10.

[23] Alferes, J., Pereira, L., Przymusinska, H., Przymusinski, T., LUPS - A Language for Updating Logic Programs, In: Proceedings of the 5th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'99), El Paso, Texas USA, December 2-4, 1999, LNAI, Vol. 1730, Springer, pp. 162-176.

[24] Berge, T., Hezewijk, R., Procedural and Declarative Knowledge. An Evolutionary Perspective, Theory & Psychology, 1999, Sage Publications, Vol. 9(5), pp. 605–624.

 ISBN: 978-1-61208-160-1

Ontology-based Foundations for Data Integration

Virginija Uzdanaviciute
Department of Information Systems
Kaunas University of Technology
Studentu street 50-313a, Kaunas
virginija.uzdanaviciute@stud.ktu.lt

Rimantas Butleris
Department of Information Systems
Kaunas University of Technology
Studentu street 50-313a, Kaunas
rimantas.butleris@ktu.lt

***Abstract*—The integration of data is one of the most complicated tasks which need to be addressed by IT researchers. Despite the critical importance, the current approaches to semantic interoperability of heterogeneous databases have not been sufficiently effective. We apply an idea of ontology as the foundation for data integration. In the paper the most common data integration methods and their essential features are discussed; the major problems of integration tasks are highlighted. Requirements for implementing data integration tasks are defined for the model, the architecture, the content, and the representation. In order to specify data source semantics, a meta-model is created, which is used to describe the concepts of source and relations. The following architectural ontology-based models are analyzed by selected criteria: a global, a multiple and a hybrid. The model of data integration process is then built upon the hybrid architecture model.**

Keywords-system interoperability; data integration methods; ontology-based data model; integration approaches; semantic integration

I. INTRODUCTION

One of the biggest current research challenges in the human–computer interaction and information retrieval is to provide users with intuitive access to the growing amount of data present in different database management systems (DBMS). Databases (DB) are designed and filled over time by different people, but they represent the same or related areas. These data express real world facts, attributes and interrelationships. The origin of data is their history which covers source collection, processing technologies, executors, etc. All businesses collect data using diverse information systems (IS). Typical ISs are designed to enable users to perform business operations and not to exchange data with other ISs. The interoperability of systems is, unfortunately, not relevant from the rate of interest standpoint. This has been a critical issue within the database community for the past two decades [27].

The interoperability of a system is seen as a consequence of technical, semantic, organizational, legal and political tools. It empowers transfer and usage of data in other information resources by following types:

- Organizational. It specifies the regulation of resource interaction.
- Technical. It describes the compatibility of IT tools, establishment and usage of open interfaces, standards and protocols in order to ensure effective data exchange.
- Semantic. This characteristic ensures that data from one IS are understood and interpreted in the same way in other systems.

Systems must be able to exchange data. Data exchange between ISs is determined by reciprocal agreements which are different in each case: web-based services, open standards, specifications [30]. Direct data integration is impossible if data is processed by applied IS logic [2]. This process can be performed in real time in source system changes occur, fixed time intervals automatically or manually, using popular methods: Extract, Transform and Load (ETL), data replication, federation, event-based integration, web-based technologies and open standards [18]. The aforementioned methods have essential disadvantages in the context of heterogeneous DBMS [15][16]: the problems of automatic update are neither considered nor solved, the same data is stored in several sources. Besides, there is no possibility to get data or information messages on databases using direct access interaction. The researchers of distributed heterogeneous databases have applied ontologies to support semantic interoperability: to integrate data sources developed using different vocabularies and to see data from a different perspective [22].

The process model presented in this paper describes the foundations for ontology-based data integration system. The described data integration task automatically performs data extraction and integration from both structured and semi-structured data sources. In addition, semantic IS interaction type is analyzed; searching for solution of ensuring unified understanding of the same data which is in heterogeneous data source systems, clearly describing semantics of commonly used data. The proposed process model integrates and reuses data using ontologies by relevant criteria.

The structure of the paper is as follows. Section 2 introduces the foundations of data integration. Section 3 relates the different concepts of the task. ER and ontology-based data models are compared; essential features of the models are highlighted. In addition, we also derive the requirements for an ontology modeling language. An ontology-based data source (OBDS) model is proposed for the development of systems. The evaluation of different aspects of the architectural models based on ontologies

 ISBN: 978-1-61208-160-1

(global, multiple and hybrid) is also presented. Furthermore, an overview of the most popular data manipulation methods is in order. Section 4 describes the process of data integration based on ontology. Section 5 presents the conclusions of our research and thoughts on future work.

II. DATA INTEGRATION

Normally, the organizational data resides in multiple data sources. For typical business intelligence (BI) data integration projects [18], the design and development of data integration processes involve collecting facts for the integration, analyzing data structures and their descriptions [4]. However, it is inappropriate to focus on the management of data requirements [2] only: it is very important to discern that integration is more than data. It also covers:

- Data sources: what data from where has to be integrated?
- Business rules (BR): which BRs have to be evaluated for data processing and keeping in data sources?
- Transformations: which transformations have to be done in order to avoid structural and semantic conflicts?

The integration of data – data management in the way that they would be unambiguously identified in IS and it is possible to transfer, transform, load and use them in other IS or source without changing program code [18]. Currently prevailing IS infrastructures are characterized as complex, distributed and heterogenic environments [1]. For this reason, data and integration of various programs are unceasing challenges for system developers, as well as providing accurate information which is necessary in today's competitive markets [18]. Ontology-guided data integration makes the process more efficient – reducing the cost, maintenance and risk of the project [18].

In order to consolidate and integrate data, we have to know which data is required, where it is and which method will ensure this process. First of all, we have to identify data, determine suitable ISs and DBMSs. Moving on, we have to specify the relationships, place, and accessibility of the data. Then we can create logical schemes, i.e. perform reverse engineering and use data dictionaries (if they exist). The next (very important) step is to evaluate DB integrity (triggers, relationships), data structure (types and lengths of fields), the time and frequency of their updates. A further step is to describe meta-data. The third step is to create the meta-data model, which describes not only data structure, but also their reciprocity. The model can be formally defined by Entity-Relationship (ER) diagram [10], which allows visual specification of data structure and relationships [17]. Alternatively we can use Data Flow Diagrams (DFD), which identify the manipulated data and visualize the data flows among processes, repositories and external environment entities [17]. Unified Modeling Language (UML) class diagrams, which describe data structures using interrelated classes, are also an option. Consider even the flexible RDF/OWL data model [15][16][23][24]. After carrying out these steps, data integration can be performed.

III. THE TASK OF DATA INTEGRATION

In this section, we describe the requirements for the task of data integration. As sketched in Figure 1, the integrated data takes into consideration the four following aspects designed in the composition model: a model, architecture, the content and the representation. Each aspect is described below:

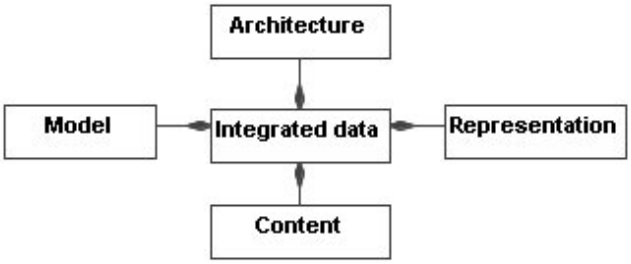

Figure 1. Integrated data.

The data model of a particular data source definition. The model must allow extension with new data, retrieval of important, highest-quality, semantically meaningful information, and the re-use of data.

The architecture, which is the core of integration and has to perform the high level autonomy to data sources. The architecture must provide semantic interoperability for the systems. It must allow the system architect to manage the development of data collections when data sources have different formats (text files, XML schemas, relational models) and to re-aggregate the application.

A neutral representation abstracts specific syntax; therefore, all the structured and semi-structured data sources first need to be expressed in a neutral format. A set of content data elements must be able to receive high-quality, semantically meaningful information. The content is heavily affected by the semantic conflict types to be resolved.

The content, i.e., the meaning of the information that is interchanged, must be understood. Data and relations have to be visualized and represented in the best, most appropriate way. It follows that each representation must bind a single expression to a single meaning using the Resource Description Framework (RDF) language.

A. *Requirements for Modeling Language*

The criteria for ontologies as modeling language of a domain modeling and formal semantic specification of data sources are chosen [5][11][22]. Minimalism is of paramount importance – only the necessary concepts must be represented in ontology. Expression – all the required concepts of a domain must be represented. Clarity – ontology must be unambiguous, easy to learn and remember; the meaning of diagrams or text expressions must be intuitively obvious, the language concepts and notations should be understandable by non-technical domain experts. Semantic stability – possibility to remain in changes of a domain. Semantic suitability – only conceptually suitable entities of a domain are modeled. Verifiability – domain experts must be able to verify if model corresponds to

 ISBN: 978-1-61208-160-1

domain. Abstract mechanisms – possibility to hide unwanted details. Formality – model is unambiguous and executable.

B. Ontology-based Data Model

The term "ontology" refers to a machine-readable representation of knowledge, particularly for automated inference. Ontology is a data model which consists of these parts: classes, properties and relationships between them [19]. The power of ontologies lies in the ability to represent relationships between the classes. The main benefit of using the ontology-based model is its runtime interpretation [21]. One of the major advantages of the ontology model is an assumption of open-world [12]. The reason for the popularity – clearly interpreted dissemination of knowledge between people and applications [7]. Moreover, ontology supports the integration task as it describes the exact content and semantics of these data sources more explicitly [1].

Gardner [18] proposes to focus explicitly on the representation of knowledge rather than just its management. He ensures that if a highly descriptive semantic representation of the available knowledge could be built, it could be reused to power a variety of business applications without the need for repeated integration exercises. Furthermore, the new knowledge gathered from different sources can build upon the current knowledge because all of it exists in a semantically consistent system. Thus we conclude that knowledge is the foundation of all successful decisions.

Semantic technologies in data integration solutions allow representing relationships of data definition area, relating data using data sets and identifying relationships for new associations. The reuse of legacy data provides the following opportunities: store and represent any types of data, easily modify the model, expand it with new data, evaluate changes.

An overview of the requirements which will be automatically satisfied by an ontology-based process is given in Figure 2.

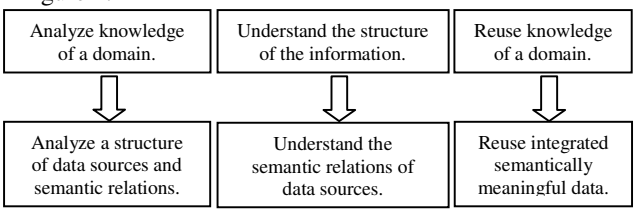

Figure 2. Transformations of ontologies features to data integration systems requirements.

Besides, ontology-based model is used to solve semantic and syntax conflicts of the heterogeneous data sources [9][22]:

- Schematic: when data is stored in heterogeneous DBMS. Such conflicts are caused by different designers, different area interpretation and usage of different data models.
- Semantic: when different class / attribute names (issues of synonyms and homonyms), different output formats (coding, data formats), different meanings are used. Conflicts arise in attributes when semantically equivalent (having the same meaning) attributes have different domains in several schemes.

C. Ontology-based data source integration architectures

In this section, we describe the three main ontology-based architectures: global, multiple, and hybrid in Figure 3 [4][7][14][20].

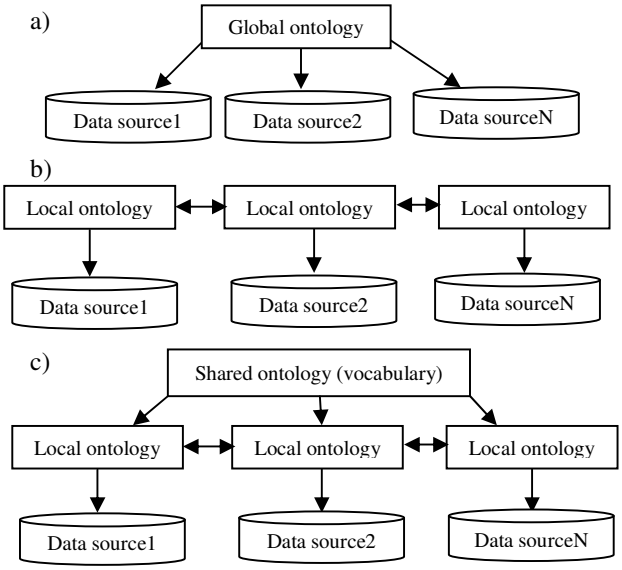

Figure 3. The methods: a) global, b) multiple, c) hybrid.

The method of data integration generally depends on which integration architecture the developer is most familiar with, what is known about heterogeneity of the data sources, etc. In models [14] shown in Figure 3 ontologies and data sources are represented as classes and semantic relationships between data source content and ontology as links, and links between local ontologies as mappings.

The global ontology method in a) of Figure 3 uses a single ontology. The method has a single main stage: building global ontology by domain expert, who knows the semantics for all data sources. The global ontology can also be a combination of several specialized ontologies. The global ontology describes data from heterogeneous data sources; query is executed via the main ontology. All data sources are related to the global ontology. This method can be applied to integration solutions where all data sources to be integrated provide the same view on a domain.

The multiple ontology method in b) of Figure 3 uses local ontologies and mapping rules between them. Each data source is described by its own ontology. The mapping rules can be modified according to the dynamic change of data source. The method has two main stages: building local ontologies and defining mappings. The local ontologies describe data from heterogeneous data sources; integrated query is executed via the local ontologies. The essential feature of this method is that the ontologies for individual data sources could be developed or changed without respect to other semantic relations, data sources or their ontologies.

The hybrid ontology method in c) of Figure 3 uses a vocabulary of a domain to represent a shared ontology, a

 ISBN: 978-1-61208-160-1

local ontology and mapping rules between them. The specification of a vocabulary includes definitions of classes, relations, functions, and other objects [8]. The mapping rules can be modified according to the dynamic change in data source. The method has three main stages: building the shared vocabulary, building local ontologies and defining the mappings. Similarly to the multiple ontology method, the semantics of each source is described by its own ontology. However, in order to make the source ontologies comparable to each other, they are built upon one global shared vocabulary.

The advantage of the hybrid method is that new data sources can easily be added without the need to modify the mappings or the shared vocabulary. Therefore, the hybrid ontology architecture gives more autonomy to data sources [28]. The use of shared vocabulary makes the source ontologies comparable and avoids the disadvantages of single or multiple ontology methods. Table 1 presents the different ontology architecture methods resulting from this analysis.

TABLE I. BENEFITS AND DRAWBACKS OF THE ONTOLOGY-BASED INTEGRATION METHODS

Criterions	Ontology-based architectures		
	Global	**Multiple**	**Hybrid**
Evaluation of semantic heterogeneity	Useful for systems which have the same view on a domain.	Useful for systems, which have the same view on a domain.	Useful for systems, which have the different view on a domain.
Appending new data sources	Some modification is necessary in the global ontology.	Supports an opportunity to append the new data source with some adaption in other ontologies.	New data sources can easily be added without the need of modification.
Elimination of data sources	Some modification is necessary in the global ontology.	Supports an opportunity to remove the data source with some adaption in other ontologies (need to remove relation between ontologies).	Data sources can easily be removed without the need of modification.
Comparison of multiple ontologies	Impossible.	Difficult, because of lack of a common vocabulary.	Simple, because ontologies use a global shared vocabulary.

D. An Ontology of Data Source

Ontology-based data integration is an effective method to cope with the heterogeneous data. This solution is based on the idea of decoupling information semantics from the data sources. Moreover, ontologies support dynamic domains better. For this reason, it is necessary to analyze data source elements: data, schema, schema elements and content, values, entities and attributes, query result classes. It is known that ontology-based search system gives the user more meaningful query results than the normal search system [13], which queries data with syntactic parameters. The query result is based on data retrieval methods [23][24][25][26]. Figure 4 gives an overview of data source meta-model.

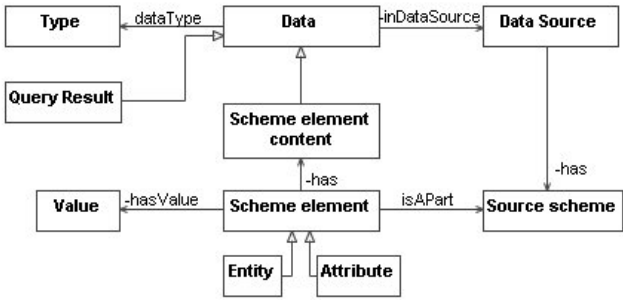

Figure 4. Meta-model of data source.

IV. CONCEPTUAL ONTOLOGY-BASED DATA INTEGRATION PROCESS

In this section, we describe the process for constructing a suitable system to semantically integrate the data from heterogeneous data sources and ensure the interoperability of it. The process model is based on hybrid method of architecture which has a shared ontology – vocabulary. Each step of the process model has its own ontology. A shared ontology is established using local ontology for each data source and a method of ontology alignment to match them. The usage of this method allow us to enhanced usability, the possibility to model mechanisms that are closer to the way we understand the real world. According to Ram *et al.* [22], ensuring interoperability of systems and knowledge-based information sharing is one of the key aspects of successful implementation.

We propose to evaluate business rules (BRs) and constraints, which ensure data integrity and correctness in the processes of data update, processing and integration. It is known that a BR is a logical statement of what to do (what actions to take) in different situations [6]. BRs can be classified by the actions in the ISs as shown in [6]. All characteristics of the data including BRs, and constraints we extract automatically from DBMS, or describe manually in a vocabulary.

Moreover, we propose to detect and solve conflicts at both data and schema levels. Han *et al.* [29] affirms that different treatment of the data structure and semantics plays a major role in IS. In this context, the scientist tries to achieve semantic coherence by eliminating semantic conflicts with a common ontology. Semantic Conflict Resolution Ontology (SCROL) provides a dynamic mechanism of comparing and manipulating contextual knowledge of each data source, which is useful in achieving semantic interoperability among heterogeneous data sources. A more detailed description of the conflict classification and

 ISBN: 978-1-61208-160-1

the method for automatic detection and resolution of various semantic conflicts in heterogeneous databases using SCROL are found in [22].

Our process model in Figure 5 is similar to the approach proposed by Skoutas *et al.* [3], which consists of five aspects. One of the chief drawbacks of that approach is inability to resolve all the semantic conflicts detections and solving processes. In addition, it does not evaluate BRs and constraints, which play the main part in the integration of data. Compared with Chang *et al.* [30], the proposed process of ontology-based data integration and analysis solves only data format conflicts and duplication.

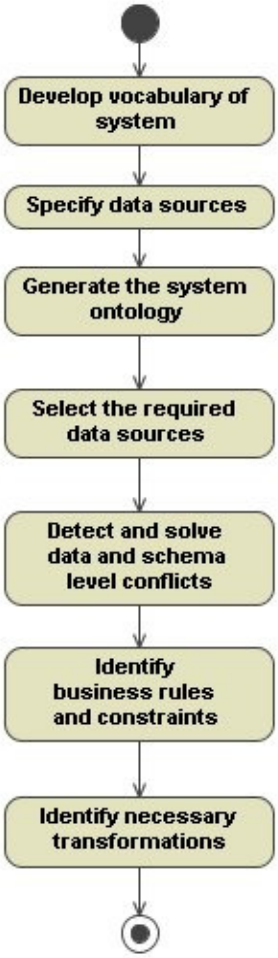

Figure 5. Process model.

The activities of our process model are specified as follows:

- *Develop vocabulary of system*. The vocabulary consists of the concepts of the domain, the attributes characterizing each concept, the different representation formats, and values for each attribute (feature values). The concepts of the domain are represented by classes, while the relationships between concepts, as well as the attributes of the concepts are represented by properties. The values of features consist of concept features and representation formats. Also, data sources and system requirements are described in the vocabulary of the system.
- *Annotate the data sources*. This one involves the mapping of attributes. Attribute mappings relate attributes to features. The types of mappings are: 'one-one', 'one-', 'many-many', and 'none'. The attribute specification consists of representation format, range of values (min, max), cardinality, referenced relations, function and attributes used for aggregation. Representation formats belong to the concept features.
- *Generate the system ontology*. System ontology is used to model the domain and to formally specify the semantics of data in the data sources. The system ontology consists of: a set of classes corresponding to the specified domain concepts, a set of properties corresponding to specified features of the concepts of the domain, and a set of classes representing the data sources.
- *Select the required data sources*. In this step we need to identify required data sources for integration.
- *Detect and solve data and schema level conflicts*. This stage is useful to determine the semantic match of data sources. It is necessary to decouple meaningful data and its semantics from the data sources with conflicting constraints.
- *Identify BRs and constraints*. It is the processing of complex BRs and constraints, including complex data transformation logic for output from integration of heterogeneous sources. Conceptual BRs and constraints provide the rationale for the correct data values in the data sources, warns of errors in the updating, processing and integrating of data. BRs ensure that the integrated data records have the same semantics. Besides, they prevent data integration between data sources with conflicting constraints and guard data correctness and integrity.
- *Identify required data conceptual transformations*. Transformation rules describe how the required data is extracted from the sources, combined and re-used in other ISs according to predefined BSs and constraints. They ensure the consolidation of data quality and detail level requirements.

V. CONCLUSION AND FUTURE WORK

We applied the idea of process based on ontology for data integration. The problem of data integration is data exchange, defined as the problem of transforming data structured under one schema into data structured under another schema.

The proposed hybrid data integration process is based on the use of ontology that explicitly captures knowledge about different types of data sources. While database schemas are generally regarded as static, the ontology schemas are typically assumed to be highly dynamic and evolving

 ISBN: 978-1-61208-160-1

objects. The key feature of the data integration process model is: it evaluates BRs, constraints and transformations for identification of semantic conflict and solving processes at both data and schema levels. The advantage of our method is that it is based on hybrid architecture method. It relies on the following elements: system vocabulary and local ontology per each heterogeneous data source. In order to integrate data from heterogeneous data sources using the hybrid method, the relations between central and local ontologies, and the relations between local ontology and the corresponding data sources should be built up.

The prospective work is to describe ontology-based data integration methodology using ontologies of the data source and resolution of semantic conflicts, BRs, and transformations.

REFERENCES

[1] M. Gagnon, "Ontology-Based Integration of Data Sources", Information Fusion, 10th International Conference, ISBN: 978-0-662-45804-3, pp. 1-8, 2007.

[2] A. Taa and A. S. Abdullah, "Norwawi M. Rameps: A Goal-Ontology Approach to Analyse the Requirements for Data Warehouse Systems", Wseas Transactions On Information Science And Applications, ISSN: 1790-0832, iss. 2, vol. 7, pp. 295-309, 2010.

[3] D. Skoutas and A. Simitsis, "Designing ETL Processes Using Semantic Web Technologies", In DOLAP'06: Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, ACM Press, pp. 67–74, 2006.

[4] Skoutas D. and A. Simitsis, "Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data", International Journal on Semantic Web and Information Systems, Special Issue on Semantic Web and Data Warehousing, vol. 3, no. 4, pp. 1-24, 2007.

[5] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford Univ., 1993.

[6] L. Tutkute and R. Butleris, "Template based business rules modelling from UML to executive code", Proceedings of BIR'2009 : the 8th International Conference on Perspectives in Business Informatics Research, Kristianstad University College, 01-02 October 2009 / J. Aidemark, S. Carlsson, B. Cronquist (Eds.), Kristianstad : Kristianstad Academic Press, 2009, ISBN 9789163355097, pp. 7-14, 2009.

[7] D. Fensel, "Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce", Springer Verlag, pp. 138, 2001.

[8] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[9] S. Spaccapietra and C. Parent, "View Integration: A step forward in solving structural conflicts", IEEE Transactions on Data and Knowledge Engineering, vol. 6, no. 2, pp. 258-274, 1994.

[10] P. McBrien and A. Poulovassilis, "A Formal Framework for ER Schema Transformation", International Conference on Conceptual Modeling, the Entity Relationship Approach, Springer, pp. 408-421, 1997.

[11] E. Vysniauskas and L. Nemuraite, "Transforming Ontology Representation from OWL to Relational Database", Information technology and control, vol. 35, no. 3A, pp. 333-343, 2006.

[12] J. Bock, S. Grimm, J. Henß and J. Kleb, "A Database Backend for OWL", In: Proceedings of the 5th International Workshop on OWL: Experiences and Directions OWLED 2009, vol. 529, pp. 1–8, 2009.

[13] Oracle Database Semantic Technologies Overview, White Paper, 2007.

[14] L. Zhang, Y. Ma and G. Wang, "An Extended Hybrid Ontology Approach to Data Integration", 2nd International Conference on Biomedical Engineering and Informatics, pp.1-4, 2009.

[15] L. Dong and H. Linpeng, "A Framework For Ontology-based Data Integration", International Conference On Internet Computing in Science and Engineering ICICSE 2008, pp. 207-214, 2008.

[16] T. Berners-Lee, "The Semantic Web", Sci. Am. 284, pp. 34-43, 2001.

[17] P. P. Chen, "The entity-relationship model – toward a unified view of data", ACM Transactions on Database Systems (TODS), vol. 1, no. 1, pp. 9-36, 1976.

[18] S. P. Gardner, "Ontologies and semantic data integration", DDT, vol. 10, no. 14, pp. 1001-1007, 2005.

[19] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Technical Report, SMI-2001-0880, Stanford Medical Informatics, Stanford, 2001.

[20] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner, "Ontology-Based Integration of Information - A Survey of Existing Approaches", In Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, pp. 108-117, 2001.

[21] D. Calvanese and G. Giacomo, "Ontology-Based Data Integration", Tutorial at the Semantic Days 2009 Conference Stavanger, Norway, 2009.

[22] S. Ram and J. Park, "Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflict", IEEE Transactions on Knowledge and Data Engineering, vol., 16, no. 2, pp. 189-202, 2004.

[23] OWL Web Ontology Language Reference. Available at: http://www.w3.org/TR/owl-ref.

[24] Resource Description Framework (RDF). Available at: http://www.w3.org/TR/rdf-concepts.

[25] SPARQL Query Language for RDF. Available at: http://www.w3.org/TR/rdf-sparql-query.

[26] XML Query. Available at: http://www.w3.org/XML/Query.

[27] A. P. Sheth, "Semantic Issues in Multidatabase Systems", SIGMOD Record, vol. 40, no. 4, pp. 5-9, 1991.

[28] L. Bellatreche, G. Pierra, "OntoAPI: An Ontology-based Data Integration Approach by an a Priori Articulation of Ontologies", 18th International Workshop on Database and Expert Systems Applications, IEEE Computer Society, pp. 799- 803, 2007.

[29] L. Han and L. Qing-zhong, "Ontology Based Resolution of Semantic Conflicts in Information Integration", Wuhan University Journal of Natural Sciences, vol. 9, no. 5, pp. 606-610, 2004.

[30] X Chang. and J. Terpenny, "Ontology-based data integration and decision support for product-Design", Robotics and Computer-Integrated Manufacturing 25, pp. 863–870, 2009.

[31] J. Heflin and J. Hendler, "Semantic Interoperability on the Web", In Proceedings of Extreme Markup Languages 2000, Graphic Communications Association, pp. 111-120, 2000.

 ISBN: 978-1-61208-160-1

Facilitating Business Process Discovery using Email Analysis

Matin Mavaddat
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Matin.Mavaddat@live.uwe.ac.uk

Stewart Green
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Stewart.Green@uwe.ac.uk

Ian Beeson
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Ian.Beeson@uwe.ac.uk

Jin Sa
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Jin.Sa@uwe.ac.uk

***Abstract*— Extracting business process models from stakeholders in large organizations is a very difficult, if not impossible, task. Many obstacles such as tacit knowledge, inaccurate descriptions of processes and miscommunication prevent process engineers from ascertaining what the business processes actually are. Data sources that represent the communications can be a good candidate for facilitating the identification of the business processes. The proposed approach in this research is to find business process related emails, identify email message threads, and finally, tag them using conversation for action theory. The outcome of this method will be process fragment enactment models that can help process engineers both to validate their findings about the business processes, and also to understand better the vague and unclear parts of the processes.**

Keywords-Business Processes; Email Analysis; Process Mining; Semantic Similarity.

I. INTRODUCTION AND BACKGROUND

"The reality is often very different from what is modelled or what people think. The world is not a Petri Net." Van der Aalst [1]. The researcher's experience confirms this, while working on two requirement extraction and business process analysis projects for Saint John Ambulance and Intellect Publishing Ltd. during the past three years. He has always spent an extensive amount of time with the clients and tried to use conventional tools and techniques, as well as best practices, for business process modelling and analysis, but there are always many obstacles in the way in order to find the actual business processes that are being followed in an organization. These are almost the same as challenges that we face in knowledge acquisition: for example, limited memory, information processing biases, representativeness, communication problems and different perceptions [8]. In this research, we are trying to find out if by using the data in the email corpus of an organization in conjunction with the conversation for action and speech act theory we can create fragments of business process enactments to help process engineers understand organizational processes better. Due to privacy issues, one of the challenges in analysing an email corpus is to have access to one. For this research the researcher has access to Intellect Publishing's email corpus. Another challenge is to what extent the business processes are being carried out using email messages. The models that are created using the proposed approach are called process instance "fragment" models and based on the amount of "fragmentation" these models can be very useful or of little use. This fragmentation is directly related to the number of business processes that are being carried out using the email system either fully or partially. Intellect Publishing, like many other organisations [3], uses emails as its main means of communication. Therefore, a good portion of its journal production processes is carried out using email messages.

The rest of the paper is organized as follows. In Section II, the main differences of the proposed approach and the existing ones will be discussed. In Section III, the solution approach will be introduced, and finally, in Section IV, we will put forward our conclusions and future work.

II. STATE OF THE ART

To date only a small amount of work has been done in this field, such as the works of Van der Aalst et al. [2] and Cohen et al. [16]. In all of these proposed works and methods some assumptions have been made that make using these techniques almost impractical. Assumptions like: Spam free mailboxes (Here spam means all the emails that are unrelated to the business processes) and finding the email threads only from email meta-data or by analysing email subjects or by manually created tags. In the proposed solution in this research, none of these assumptions have been made. Email messages are automatically filtered using text categorisation techniques and analysing emails' content semantics is added to the subject and meta-data analysis in order to overcome the shortcomings of previously introduced

 ISBN: 978-1-61208-160-1

methods. In addition to the aforementioned points, none of the previous methods have used the conversation for action and speech act theory to justify the soundness of this idea that email analysis might be a good method for facilitating the business process identification.

III. SOLUTION APPROACH

The final goal of this research is to create a method both to extract business process related emails from an email corpus, and to create a model from them that shows the interaction between different role instances involved in the business process. This method may help the process engineers validate or better understand the processes that are elicited by other means. The method has three main stages: email categorisation, conversation network finding, and conversation network tagging.

A. Email categorization

The first process, which feeds directly from the email corpus, is the email categorization process. In email categorization, text classification techniques are used. Text classification is assigning automatically a text document to a predefined class [8]. Obviously each email corpus contains several different types of emails such as business-rule related emails and personal emails which might not contain any business process related information. So it is necessary that the email corpus be divided (classified) into two different classes: business process related and non-business process related.

Many different methods and techniques have been introduced to solve this type of problem such as Naïve Bayes [7] and Support Vector Machines [5]. The main challenge in solving this problem is to find the features that help to classify textual documents (emails) into business process related and non-business process related.

In order to use automatic categorization techniques and also in order to choose the best categorization algorithm, a number of email messages should be classified manually to create a training set and a test set. A text-mining tool named WEKA [13] can be used to test different text mining algorithms. The training set can be fed to the WEKA tool, which automatically trains itself using different text mining training algorithms, including the ones that were mentioned before. WEKA automatically extracts classifying features that distinguish business process related emails from non-business process related emails using different algorithms, then, using the test set and feeding it to the tool, the algorithm that best matches the human classification of the emails can be identified.

B. Conversation network finder

After classifying the emails and finding the business process related emails, the next step involves finding email threads inside the business process related emails. Email threads, or email trees, are related email conversations that have occurred about a similar topic, ordered by time. Email threads that are created from the classified set of emails can be a representation of a network of "conversation for action" introduced by Winograd [12]. Winograd put forth the conversation for action theory based on Searl's [10] speech act theory. He argues that business processes are networks of conversations that are happening inside and outside the organization about the organizational goals. They introduce the conversation for action diagram and believe that almost every network of conversation for action happens according to the pattern introduced in that diagram. He believes "speech acts are not individual unrelated events, but participate in large conversational structure." [13]

For example, one conversation for action network might start with a request, and a request is understood by the participant as having certain conditions of satisfaction, which shows how the hearer might react to the speaker's request. Finding these networks of conversation for action" should help us realize organization's business processes.

 ISBN: 978-1-61208-160-1

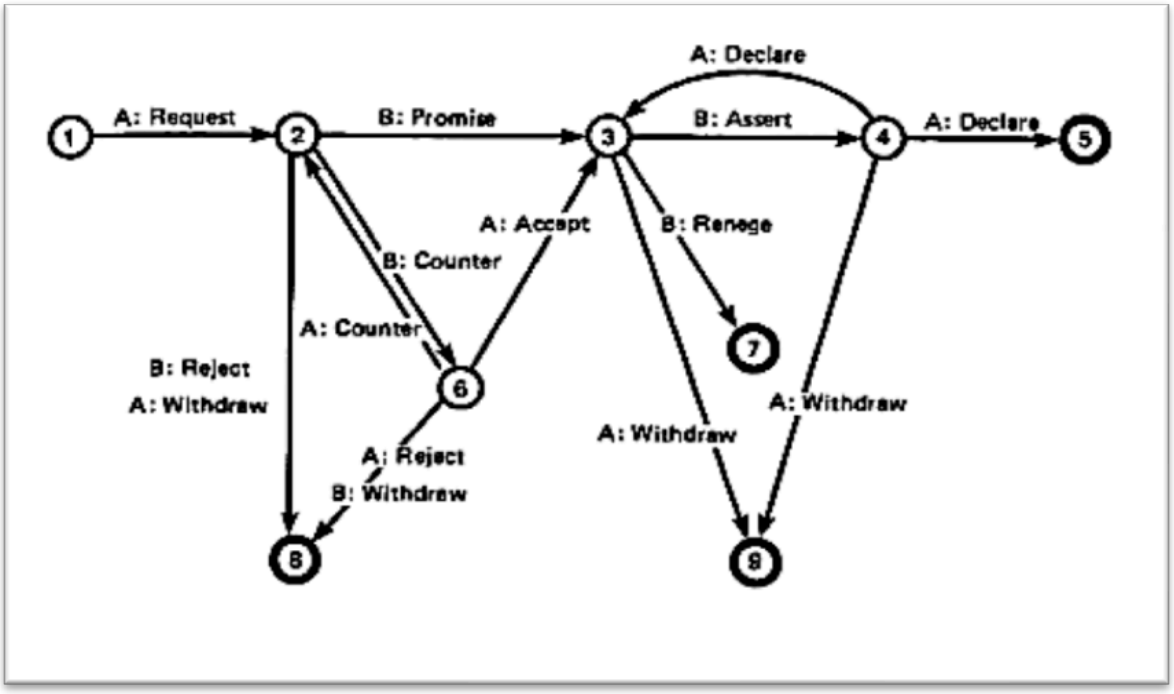

Figure 1. Conversation for Action diagram

One approach to find these networks of conversation is to use semantic similarity and some other heuristics [15]. By applying this process, the business process related set of emails turns into different threads, each representing one conversation network that is initiated by a role instance and has been continued until a mutual agreement.

One problem here is that if all the conversions have not happened via email, then it cannot be realized. This is the problem of incompleteness in process mining. This means we may not be able to model the whole process. This is why it is called a process instance (enactment) "fragment" extraction.

A Java application has been developed for this purpose and a function within this application is created to measure the semantic similarity of two email messages in order to realize if they are related. This is used in email thread finding. It is assumed that if an email is sent by a recipient of an original email after the date of the original email, and it is semantically similar to the original email, then it is part of the thread that is initiated by the original email.

A refined version of Vector Space Model algorithm [7] is used in this function. It means that each email is translated into a vector implemented in sparse matrix and the semantic similarity is measured using the multiplication of vectors.

The following refinements have been made to the original vector space model: first, the vectors are not two-dimensional but n-dimensional. The synonyms are added to the vector as the higher dimensions using WordNet [4]. It is possible in future to just add the context-related synonyms (by interpreting and using the WordNet as a lexical ontology and comparing it with the existing ontologies in the organisation); and second, Checking the words' spellings before adding them to the vector using spell checking algorithms.

The output of this function will be something like the following table for the simplified scenario below:

1. Matin sends an email to Stewart to request for a meeting and specifies his availability.
2. Stewart selects a date and time and sends an email to Jin to ask if she can make it too.
3. Jin sends an email to Stewart and accepts the date and time.
4. Stewart sends an email to Ian to see if he can make it at that date and time.
5. Ian responds that he can make it.
6. Stewart sends an email to Matin telling him the date and time of the meeting.

TABLE I. CONVESATION NETWORK FINDER OUTPUT TABLE

#	**Sender**	**Recipient**	**Timestamp**	**Email ID**
1	matin @uwe.ac.uk	Stewart@uwe.ac.uk	15-Jun-2010 14:00:00	132
2	Stewart @uwe.ac.uk	Jin@uwe.ac.uk	15-Jun-2010	139

 ISBN: 978-1-61208-160-1

			14:40:23	
3	Jin @uwe.ac.uk	Stewart@uwe.ac.uk	15-Jun-2010 15:39:30	150
4	Stewart @uwe.ac.uk	Ian@uwe.ac.uk	16-Jun-2010 09:30:30	180
5	Ian @uwe.ac.uk	Stewart@uwe.ac.uk	16-Jun-2010 12:30:12	200
6	Stewart @uwe.ac.uk	Matin@uwe.ac.uk	16-Jun-201015:30	250

The results created by this application were quite satisfactory when they were analyzed manually. The extracted threads were quite reasonable and close to the actual threads that were created manually.

C. *Conversation network tagging*

Up to this stage, the relation between role instances have been extracted; this shows the interaction between different roles instances to achieve a goal (mutual satisfaction), but the interactions are not labelled. This labelling can be done using the speech act theory [11] as this theory tries to define what the speaker intends to do by using words.

Searle [10] has set up the following classification of illocutionary speech acts: assertives, directives, commissives, expressives, and declarations.

By finding the illocutionary acts and the propositional content related to each email (or email paragraph if needed), we can label each interaction. For example, the result of this step is expressed by the following RAD-like diagram [9]:

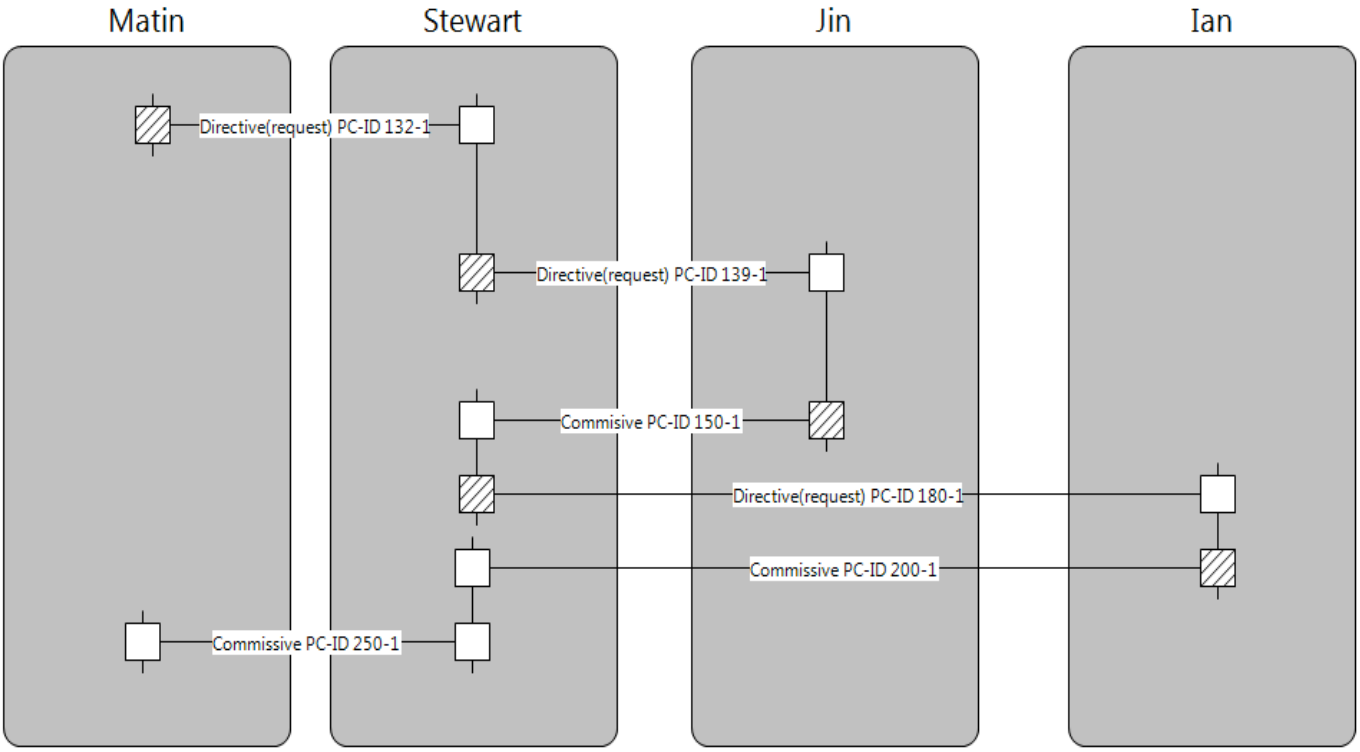

Figure 2. Conversation networks.

Fig. 2 shows the interactions between different roles instances and the interactions are tagged by the extracted illocutionary act of the relevant paragraph or email and also a link to the actual paragraph or email for manual reference and further analysis.

This model is an interaction fragment instance. By analysing these models and finding the similar patterns, we should be able to create business process fragment instance models. For instance, by analysing the interaction fragment instance models of different meeting scheduling occurrences and finding the patterns, we should be able to create meeting scheduling business process fragment models. These models may help business engineers to validate their understanding of the business processes, and also might clarify some vague fragments of the manually identified processes. Interpreting these models should not be difficult for the process engineers. RAD diagram notations are being used, the interactions are tagged by illocutionary acts that are quite

 ISBN: 978-1-61208-160-1

self-explanatory and each interaction is directly linked to the actual paragraph or email for further analysis if something is not clear enough.

IV. CONCLUSION AND FUTURE WORK

This paper proposed an approach to facilitate the identification of business processes of an organization Fig. 3. For the first sub process a database is created from the emails metadata and content, for analysis with WEKA for email categorization and for the second sub-process an application has been developed that identifies the threads using a modified version of the Vector Space Model algorithm. The third sub-process needs more research involving both tagging identified conversations using the "speech act theory", and using the "conversation for action" idea to find similar patterns and to create the fragments of business process instances from the interaction fragment instances.

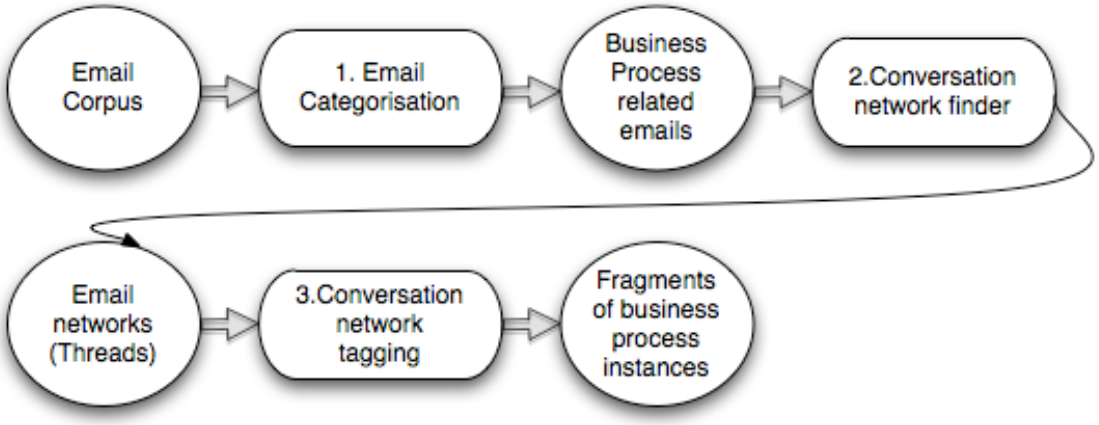

Figure 3. Solution Approach

REFERENCES

[1] Aalst, W. M. P., 2003. Challenges in business process analysis. Bulletin of the EATCS, 80, pp. 174-198.

[2] Aalst, W. M. P. and Nikolov, A., 2008. Mining e-mail messages: uncovering interaction patterns and processes using e-mail logs. International Journal of Intelligent Information Technologies, 4 (3), pp. 27-45.

[3] Ellis, C. A., 2000. An evaluation framework for collaborative systems, University of Colorado at Boulder, USA.

[4] Fellbaum, C. 1998. WordNet: An electronic lexical database (Language, Speech and Communication), MIT Press.

[5] Joachims, T. , 1998. Lecture Notes in Computer Science-Text categorization with Support Vector Machines: Learning with many relevant features, Springer, pp. 137-142.

[6] Liu, B., Lee, W. S., Yu, P. S., and Li, X., 2002. Partially supervised classification of text documents. ed. 19th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 387 – 394.

[7] Manning, C. D., Raghavan P., and Schutz H., 2008. An introduction to Information Retrieval. Cambridge University Press.

[8] Mcgraw, K. L. and Harbison-Briggs, K., 1989. Knowledge acquisition principles and guidelines, Prentice- Hall.

[9] Ould, M., 2005. Business Process Management: A Rigorous Approach, BCS.

[10] Searle, J. R., 1975. A Taxonomy of Illocutionary Acts, in: Günderson, K. (ed.), Language, Mind, and Knowledge, Minneapolis, vol. 7, pp. 1-29.

[11] Searle , J. R., 1969. Speach Acts: An Essay in the Philosophy of Language Cambridge University Press.

[12] Winograd T. and Flores, F. , 1986. Understanding computers and cognition, Addison-Wesley.

[13] Winograd, T. , 1987. A language/action perspective on the design of cooperative work. Human-Computer Interaction, 3 (1), pp. 3-30.

[14] Witten, H. I. and Frank, E. , 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann publishing.

[15] Yeh, J. and Harnly, A, "Email thread reassembly using similarity matching." Proc. Third Conference on Email and Anti-Spam (CEAS), 2006.

[16] Cohen, W., Carvalho, V., and Mitchell, T., 2004. Learning to classify email into "Speech Acts.". Association for Computational Linguistics,4(11), pp. 309-316, doi: 10.1002/asi.20427.

 ISBN: 978-1-61208-160-1

Reusable Decision Models Supporting Organizational Design in Business Process Management

Olga Levina, Oliver Holschke
Berlin Institute of Technology
Department of Systems Analysis and IT
10587 Berlin, Germany
{olga.levina; oliver.holschke}@sysedv.tu-berlin.de

***Abstract*—Transition towards a business process oriented enterprise is explored in research from different perspectives. One of the most popular approaches is Business Process Management. Its activities include process definition, execution, monitoring and analysis. Its implementation was studied in research and practice from the process and application related points of view. The organizational changes are often out of scope or only marginally regarded in these implementation strategies. This paper summarizes recurring and thus reusable organizational decisions that appear in the course of the implementation process and provides the accordant organizational design. This design can be applied throughout the industries for enterprises that are about to start a business process management initiative. Thus, this approach supports a rather inter-enterprise than intra-enterprise reuse paradigms.**

***Keywords*-organizational patterns; business process management; design model reuse.**

I. INTRODUCTION

There are numerous Business Process Management (BPM) implementation strategies, mainly provided by large consultancy firms or business application software developers that individually emphasize different aspects such as general implementation strategies as well as choosing and implementing a business process modeling tool. Among these, the organizational changes are very important as the organization commits to implementing business process management activities [25][29][22]. The organizational structure is an essential attribute of the design of an enterprise, especially when in a situation of change (e.g., BPM implementation). It defines the relationships between employees to support specialization and coordination of work. Modern enterprises have a variety of relationships between organizational units in order to operate more efficiently and responsively [20]. In case of an enterprise orienting towards a process-centric views and BPM the allocation of activities to positions and establishing new communications structures between them is a challenge in itself – this being decided often before thought is given to what detailed activities will be executed and what exact IT will support this. Therefore, managers need to be able to develop robust models of their organizations regarding the BPM-related changes, and then provide them so the information can be effectively reused and shared by many people and systems involved in the BPM project. Although organizational changes are usually determined early in the process of emerging BPM commitment – particularly when BPM should pervade several divisions or even the whole enterprise – actionable artifacts that may support the manager in allocating actions and responsibilities to organizational entities are scarce.

Artifacts of today do not meet the requirements of decision makers that need to be guided in the organizational design of BPM implementation projects. However, the availability of such a decision supporting artifact and its reuse across different BPM initiatives may greatly benefit organizations that have adopted business process-centric views. These benefits can be found in a guided and faster way of identifying and making organizational BPM decisions and reduced risk of failing to specify organizational-critical parameters which may have led to uncalled-for costs [28]. Reusing such knowledge on organizational BPM decision making within a globally operating enterprise and its distributed business units would also be a prerequisite for future innovative behavior [12]. Acknowledging that different organizations display a wide range of characteristics which will require individual strategies contingent upon the company's composites and its environment, we argue that recurring decision topics in organizational design do exist and that they may be consolidated and "packaged" for later reuse in similar problem settings [13]. We therefore explore the possibility of creating a generalized conceptual decision model that consolidates recurring decision topics concerning organizational design in BPM initiatives. This decision model is intended as an actionable item and for later reuse by decision makers or managers in BPM projects.

The rest of the paper is organized as follows. In Section II we review the state of the art of decision support methods regarding organizational design for BPM and highlight the lack of an artifact at the adequate level. Section III collects recurring decision topics in organizational BPM design and explains their range of implementations alternatives. In Section IV we consolidate our findings and propose the developed conceptual decision model. In Section V we apply our decision model to a simplified real-world BPM implementation scenario and demonstrate the instantiation. Section VI discusses briefly the limitations of the decision model and its possible extensions.

 ISBN: 978-1-61208-160-1

II. State of the art

BPM includes methods, techniques, and tools to support the activities of design, enactment, management, and analysis of operational business processes [26]. These main activities and the scope of BPM are often summarized in a BPM life-cycle [23][17][30]. Over the last three decades research in BPM has generated a large body of artifacts addressing organizational, managerial, information systems and technology, and socio-economic topics. Works on BPM implementation often focus on the areas of business process automation or product implementation, e.g., [25][11][16]. The human aspect is considered in the focus of change management or software requirements in some of these works [25][5]. As we concentrate on the organizational design for BPM, implementation strategies regarding technology selection and process modeling as well as modelling-related topics are out of our scope. These areas are already covered by existing approaches. In the area of tool selection ISO [8] offers a standard process for software evaluation, other frameworks exist in research and practice e.g., [2][4][27]. Zucchi and Edwards [29] as well as numerous researchers such as [1] identify critical success factors for BPM implementation. In e.g. [27][9], a general process for BPM implementation and integration is shown and described. Armistead and Machin [1] propose a decision support method for selecting the adequate BPM implementation project type.

Four possible BPM implementation project scenarios are suggested that are contingent upon a) the commitment of the responsible business manager and b) the resulting impact on the organization. By determining these contingencies the fitting project type can be identified. The project types are further characterized by six features, of which two address distinct organizational attributes, i.e.: 1) the number of employees impacted by the BPM activities, and 2) the type/size of organizational unit concerned (e.g., department – business unit – organization-wide). While the decision matrix can support the broad selection of an initial BPM approach, the level of analysis remains rather high considering that organizational change for BPM will involve more operational decisions. Notably, concepts that capture coordination- as well as strategy-realted topics are not provided.

Organizational aspects of BPM, i.e., the reporting and steering department for BPM, were investigated by Gartner in multiple case studies [18][3][19]. On the alignment of the BPM Gartner states that the optimum implementation is often a blended model in which IT and the business drive BPM projects. Thus, it is suitable to align BPM initiatives and expert knowledge at the IT and gradually hand it over to the business. But first of all the common ground should be established that indicates what is the purpose and goal of BPM implementation and what results are expected. Communication and transparency on these objectives need to be provided by the higher executive levels. The lack of suitable communication structures between top executives and the business units represents another significant risk. It should be clear that the existing, often function-oriented, organizational structure will be changed towards a process-oriented organization form. These changes and their effects need to be communicated using the techniques known from the change management domain. Snabe [24] provide an extensive roadmap for enterprises to enhance their business process management performance. Nevertheless, it does not consider the assignment of responsibilities to roles involved in business process management initiation and support. Additionally, no description on governance structures for BPM are given, thus, the focus is on the processes. The considered strategy aspects focus on enterprise with a level of BPM implementation that is comparable to the CMM level 2, instead of considering level 1 that includes the decision and introduction of a governance structure. For implications of strategy definition, the extensive work by Mintzberg [15] is referred to.

One of the important BPM implementation tools is the establishment of a central expertise institution for BPM. Jeston and Nelis [9] and Olding and Hill [19] provide some approaches concerning design and implementation of such a central institution. This institution is often referred to as: “competency center” [3] or “center of excellence” [6]. Establishing such an organ requires significant organizational changes such as its alignment, its internal organization, its target group and its main tasks. Zucchi and Edwards [29] describe these institutions as “business process offices” and indicates their establishment as one of the critical success factors for BPM. Remarkably, most of the publications concerned with this topic are often case-specific or derive from consultancy practice and do not address the organizational design in their demonstrations.

III. Recurring Decisions in BPM Organizational Design

In BPM literature and in practice several implementation relevant aspects regarding organizational design exist. Here they are called *organizational decision topics*. We define organizational decision topic as a focus area including multiple potential outcomes related to an element of the organization, i.e., a managerial position, in terms of responsibility and control. Thus, a decision topic directly relates to a concrete role and scope of responsibilities in the area of interest. Each decision topic has different entities or action alternatives.

Decision on the choice between these elements is assigned to the responsible that is associated to the decision topic. We define *organizational decision* as a choice of one of the alternative entities within the decision topic. Therefore, it is a reaction to an identified problem within a decision topic, which solution provides an added value to an organizational stakeholder. We emphasize that the organizational decisions discussed here are all executive decisions following the taxonomy of Kruchten [10], i.e., they do not address requirements for technology instantiations or for their properties (cf. executive, ban and property decisions in [10]). Zimmermann et al. [28] provide three stages of the decision making in the context of decision modeling with reuse: decision identification, decision making and decision enforcement. In this paper we focus on the first two steps in the context of organizational decisions for BPM implementation. Recurring and thus reusable organizational decision topics in the context of BPM were identified

 ISBN: 978-1-61208-160-1

within BPM implementation literature and in interviews lead with BPM experts and managers that were concerned with the implementation of a BPM initiative. These topics were also verified in workshops. The interview analysis revealed organization-related content clusters that occurred in the most of the interviews. These clusters were identified as decision relevant topics. The results were then presented to a group of BPM experts including the interviewees. In the upcoming discussion, the topics were rounded up and commented. As a result of this workshop the organizational decision topics presented in figure 1 were defined. For a better representation the recurring decision topics are summarized here as questions that can occur in different order and influence one another during the implementation process. Therefore, figure 1 shows the questions that are to be considered for BPM organizational design as a cycle and not as a process model.

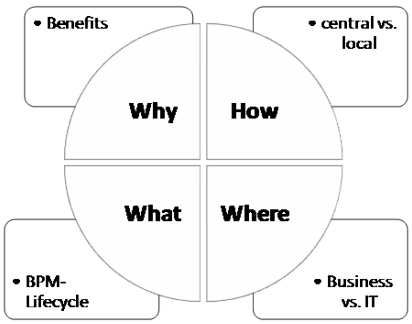

Figure 1. Reusable decisions Topics in BPM Organizational Design

BPM is a holistic management approach aiming to align an organization with the interests of its stakeholders, customers and employees while eventually raising a company‘s agility and operational performance. BPM fosters and supports business effectiveness and efficiency, adopts a cross-departmental approach and examines the impact of all relevant applications, users and stakeholders. As BPM strives for innovation, flexibility and integration with technology through continuous process improvement and the visualization of formerly non-transparent process structures, BPM is claimed to be able to increase efficiency, to reduce costs and improve the quality of products and services [9]. These arguments can serve as possible answers to the question “why” that is often present when a new technology, management approach or product program are introduced or their implementation is described. Thus, the motivation for BPM implementation can be extrinsic, i.e., enforced by law or market, or strategic. Often the question “why” BPM is to be implemented answered by general BPM strategic benefits, e.g.: increasing customer satisfaction, increasing competitiveness or more agile business process [14].

Once the BPM implementation has been decided the mode of its introduction and integration into the enterprise architecture, i.e., “how”, has to be defined. Coordination and governance aspects need to be considered within the enterprise. Governance aspects include decisions that can range from self-government to strict control. Furthermore, reporting structures need to be taken into account. Here two major modes can be identified: centralized implementation and administration or creation of local entities that are responsible for BPM implementation and expertise. A centralized BPM institution has the advantage of concentrated expertise and serving as an information point but it also requires the ability to address distributed requests. If BPM is to be introduced as a central corporate issue, it should be steered either by a central committee or a center of excellence. This central organ can be organized in different structures which are described in the following:

- Line organization,
- Staff organization,
- Matrix organization or
- Pure project organization

Which style the institution is to be organized in, is determined by the structure of the existing organization, the size, the significance and the length of the project, problem relevance and availability of material and human resources. The line organization uses the existing functional organizational departments and draws the project team members from existent business units. Therefore, one major advantage of this organization form is that the existing organization structure can be used and that team members can also work part-time for the project not hindering their normal work. On the other hand, there is the staff organization form which also uses member of a higher management level who coordinates the work of the sub project departments. In contrast, a pure project organization is most suitable if very important and huge projects are to be conducted which are not time crucial and whose budget is not too limited. This type creates a kind of organization unit just for project purposes provided with its own material and personal resources.

Finally, a project can be organized in a matrix style which follows the idea that each project member remains in his or her job position but is now subordinated to the (external) project manager who has the functional authority. In some way, the matrix organization comprises the line- and the project organization. Still, this form makes it reasonably difficult to coordinate the competency interfaces. In a hierarchical organization structure, the entities responsible for the BPM implementation may be the corporation, a division, a department, a group or a team. In addition, it may be a committee, a task force, a project management organization or any other form of a sub-team. Thus, “how” addresses distribution of authority. As modern organizations may not be confined to a strict hierarchy for chains of authority, different chains of authority may be defined for different purposes. BPM is supported by the IT (through business process execution software tools or tools for process modeling support), but it is also used in the day-to-day work by the business while IT is often needed to implement some of the BPM activities.

So the question arises, “where” BPM should be situated within the organization and what reporting structure will be appropriate. BPM responsibilities can either be situated in the IT or the business sector. Often, BPM executives report to the IT department since BPM highly depends on methodologies, leads to thorough process analysis and is supported by technologies such as modeling tools [19]. Despite the strengths

 ISBN: 978-1-61208-160-1

of IT to provide methodologies, tools and technologies, this approach leads to a lack of alignment with business objectives. Hence, reporting to the business sector is usually the better solution to ensure proper alignment with the core strategy and the company's objectives, allowing the business to better direct and manage the activities connected with economic issues [19]. When determining the relevance of the two company sectors the BPM strategy must be considered. The alignment to the IT or business sector should be determined according to the BPM strategy and employed instruments. If BPM is heading towards process automation or a BPM software tool is applied, IT must be included to some extent but should not assume the majority of BPM in the long term. Thus, "where" addresses the type and scope of organizational units involved. Reporting and communication relationships need to be established between the participating organizational entities. "Where" may occur to be the most extensive organizational decision topics, as it contains different aspects such as reporting structure, responsibility and the scope of the action. The choice depends here on multiple aspects such as the alignment of BPM stakeholders, budget responsibilities, organization structure or the dominance of the one of the two enterprise departments (business vs. IT) within the organization structure.

Main BPM activities are summarized in the so called BPM life-cycle. In literature numerous variations of the BPM life-cycle exist, its main activities can be summarized as being: Modeling, analysis, simulation, implementation, monitoring and improvement of business processes [26][30][5]. These are also the potential answers to the question, "what" should be implemented as BPM main activities within the given enterprise. Process understanding is crucial for process management that is why the most BPM projects start with introduction of a modeling tool. Process implementation is often realized using workflow management systems. Process analysis can be supported by software tools like decision support systems or business intelligence tools but it still requires significant human involvement. Establishing a process monitoring strategy is gaining more and more attention from practitioners and research. The need for a central and holistic implementation of BPM can therefore originate from the use or need to perform at least one of these activities in a business unit. Successful implementation of one of the aspects can lead to the initialization of an enterprise-wide BPM-implementation discussion, i.e., influencing the question "why".

"What" also implicitly addresses the organizational positions, i.e. extending the term of a "role", and assignments. People are linked to positions through assignments. The possible positions of interest here arise from the typical BPM activities such as designing, enacting, etc. [26][30]. These activities are often very closely related with the process execution and are often supervised and executed by the accordant business worker who is involved into the process. Though, the addressee of the results might be on the other organizational level.

IV. Conceptual Decision Model

an organization has to face in the context of a BPM implementation project is the first step towards adaptation of the organizational structures. The questions presented above are the identified as recurring decision topics during a BPM implementation initiative and are now grouped. Once the decision topics are identified, one of the alternative decision entities has to be chosen, i.e., a decision needs to be made. Figure 2 aligns decision makers with their organizational level and the decision topics presented above. Thus, the choice between the decision entities is to be made on the assigned organizational level. BPM is a topic that is relevant for the entire enterprise and is supposed to support strategic goals of the enterprise. That is why the decision on its potential benefits and implementation scope is to be supported by and taken on the upper organizational level, i.e., the so called C-level.

The "why" can also be initiated on the operational level, as the need for process modeling or monitoring can originate closer to the process execution. Decisions made on this level have a rather long-term time horizon for their realization, being mostly abstract in their declarations. The more it comes to the actual implementation and operationalization of BPM, the nearer the decision making level is to the specific business processes. The implementation of BPM expertise in the organizational structure, as well as its resulting governance and reporting structures are defined on the tactical, i.e., mid-term, level. It is often also the task of the middle-management to find a respective structure for BPM implementation. A centralized institution such as a center of excellence for BPM can often be a valid solution for big and middle enterprises. Hence, the strategy and internal organization structure need to be developed before the implementation project. Often, members of such a center are assigned to their duties despite their daily job using the staff or line organization. This structure might hinder the development of the center or the fulfillment of its tasks as the members get less identified with BPM being responsible for the daily as well as for the BPM related tasks. Furthermore, the question of whether the BPM- related topics and projects are initiated and governed, by the IT or by the business, needs to be defined on the middle-management level.

What BPM activities are needed for the actual processes performance in the enterprise can be decided or at least initiated on the operational level. That is, where the requirements are made on the process task level and the changes can be directly monitored and controlled. Additionally, tactical level can initiate the choice by requiring certain process- related quality standards or metrics. Therefore, the addressees of the BPM activity results are the operational, directly process-related workers, as well as the middle-management.

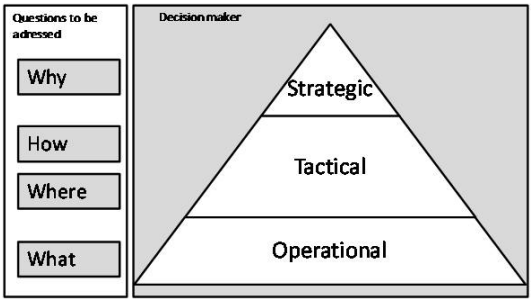

Figure 2. Conceptual Decision Model for organizational design for BPM

 ISBN: 978-1-61208-160-1

The decision structure presented in figure 2 identifies possible and reusable decisions and the accordant decision makers. This structure does not prescribe the timely sequence of the decision making. Moreover, often the decision occurrence depends on the organizational and process structure as well as culture within the enterprise. The need and benefits of BPM or some of its areas of activity are detected on the tactical or operational level. So the "what" decision often drives the "why". Was the need for BPM detected or more important supported in the IT-division, it is very likely that the BPM competences and governance structure will be tied to the IT. While the recurring decisions can influence one another in their outcomes, it is important that the roles that are responsible for their making are assigned and communicated. Addressing the main BPM topics and a general organizational structure, the decision model shown in figure 2 is not limited in use within one enterprise. Small, middle and big enterprises independent from their industrial sector can benefit from this decision structure. The reuse focus lies here on using the model for different (kinds) of enterprises in the same context: initiating or implementing business process management in their business and organizational structure.

V. Application of the Decision Model

A short illustrative example of the application of the presented decision model was conducted in a large organization that was about to develop a common BPM strategy as some BPM initiatives were already emerging developing their own approaches and tools. The strongly simplified example is used here to introduce a possible scenatio for when and where the descirbed BPM decisions can be applied including not only the business, staretgic but also the governance aspects of such an initiative.

The need for BPM in the observed company came from the operations, i.e., from the business side, decision sequence running from the bottom to the top. The enterprise has a rather flat organizational structure. The names were simplified here, as the case study is only used for demonstration purposes. Mr. A is the CEO of the considered enterprise; Mrs. B is head of the business department, while Mr. C is head of the IT department. Table 1 shows the organizational decision topics, the chosen decision outcome, decision implementation and the (made anonymous) decision maker.

Thus, the company decided to implement BPM as previous experience in business process modeling and formulization in the departments already existed. Some of the departments were already using BPM modeling tools or process management approaches. Another motivation was that BPM goals are in this case accordant with the most of the enterprise's strategic goals, i.e., design of flexible processes, short time-to-market and increasing customer satisfaction. What strategic goals can be supported by BPM implementation and how was described and operationalized in a strategy paper. The BPM implementation was decided to be central, i.e., realized by a central institution, here referred to as business process office (BPO). Matrix style organization form was chosen for the BPO, i.e., employees in the BPO still had their tasks in the departments but the tasks were limited in favor of tasks related to the BPO. Head of the BPO is Mrs. B, i.e., the business side. Tasks of the BPO were defined as: harmonization of already existing BPM initiatives, development of a general BPM strategy, IT and consulting support in BPM related fields, monitoring and control of BPM implementation.

Organizational Decision Topic	Decision	Person	Decision Implementation
Why	Support strategic goals	Mr. A	Strategy paper
How	Central	Mrs. B Mrs. B	Business Process Office Matrix
Where	Business and IT	Mrs. B and Mr. C	Business and IT
What	Monitoring, Modeling, Automation, Tool Support	Mrs. B and Mr. C	Monitoring strategy, common modeling notation, company-wide tool support

Table I
Application of the decision model

Though the initiative for BPM came from the business side, one important step for its realization was the acquisition of BPM tools, like business process modeling and simulation tools. Thus, the responsibility and therefore the budget were assigned to the IT department. Within the BPO the governance is therefore being shared by both the IT and business; IT-department being responsible for the budget and tool support while business department suggests the strategy and provides BPM expertise. Reporting on the success of BPM-related activities and projects includes the business side, i.e., Mrs. B as the head of BPO and Mr. C as head of the IT department. Once the organizational backbone has been established, the BPO took up its projects. The finished BPO activities in the company by now include: company-wide tool support, definition of a common business process modeling notation as well as developing a business process monitoring strategy. Decision on these strategies were made by Mrs. B and Mr. C based on BPM knowledge and experience as well as on the survey among the employees concerned with the topic of BPM or already having experience with BPM activities respectively.

Thus, the reusable model was helpful in this example by defining the governance aspects of the BPM initiative in the observed enterprise. A central institution for BPM was established, tasks and organizational aspects of the institution as well as the reporting structures were defined. The application of the reusable model enabled an efficient proceeding in defining, deciding and establishing the BPM supporting structures in the enterprise. Though it is too early to define the improvements of the company's process management originating in the proposed structure, the company is being observed and interviews with the BPO members are held regularly for further analysis. Additionally, the exemplary case study only illustrates the situation in one given enterprise and marks the effectiveness of the reuse model comparing to the former status quo in the company, that is multiple, scattered BPM

 ISBN: 978-1-61208-160-1

initiatives. Nevertheless, further informal interviews were lead with different enterprises that lead to the conclusion that the topics addressed in the reusable model are the areas that require coordination and managerial effort. These aspects will be the focus of our future research.

VI. Conclusion and Outlook

In this paper we addressed the fact that in the context of the implementation of business process management or process-orientation several organizational decisions have to be made. We argued that there are specific related and reusable content clusters in the area of the organizational structure that reoccur independently from industries or enterprise size during such projects. These organizational decision topics are rarely the center of attention during BPM implementation and are often being solved intuitively accordant to the ad-hoc situation. This treatment might lead to conflict situation or lack of decision power in the future. According to practical experience and literature research we provided a structured overview of these recurring and reusable decision topics as well as their entities and developed a model for organizational decision making aligning organizational decision topics with organizational roles. This approach is similar to the existing approaches in the domains of strategy definition and strategic management (as in e.g., [15]). Being focused on initial and organizational aspects of BPM implementation in the enterprise, other BPM-related aspects that occur in later stages like process modeling techniques and notations (see e.g., [21]) as well as model ontologies (see e.g., [7]) are not considered in this paper.

Literature research as well as informal interviews with business process management consultants indicated the possibility as well as supported the assumption that there are topics that recur in the scenarios of business process management initiation. These topics were summarized here to a reusable framework that can be applied in organizations that are about to introduce the BPM approach. This framework can be applied and therefore reused for decision making independently from the target industry or enterprise type in the context of BPM implementation. Next steps can include development of a method framework for decision support as well as deeper exploration of the different entities of the identified organizational decision topics.

References

[1] C. Armistead and S. Machin. Implications of business process management for operations management. *International Journal of Operations and Production Management,*, 17(9):886–898, 1997.

[2] J. Becker, M. Kugeler, and M. Rosemann. *Process management: a guide for the design of business processes*, volume ISBN-10: 9783540434993. Springer, 2003.

[3] Cognos. Building a business intelligence competency center. *Cognos Incorporated, Burlington, USA*, http://www.super-info.com.cn/innovationcenter/pdfs,(June, 2011), 2008.

[4] B. Curtis, M.I. Kellner, and J. Over. Process modeling. *Communication of ACM*, 35:75–90, 1992.

[5] D. J. Elzinga, T. Horak, C.-Y. Lee, and C. Bruner. Business process management: Survey and methodology. *IEEE Transactions on Engineering Management*, 42(2):119 – 128, 1995.

[6] Gartner. Gartner's top 10 strategic technologies for 2008. *http://www.gartner.de/fokus/07104210.pdf*, (May, 2010), 2007.

[7] Oren E. Haller, A. and P. Kotinurmi. An ontology for internal and external business processes. In *Proceedings of the 15th international conference on World Wide Web*, pages 1055–1056, 2006.

[8] ISO. Iso/iec 14598-5. software engineering - product evaluation. page http://www.iso.org/iso, 1998.

[9] J. Jeston and J. Nelis. *Business Process Management*. Elsevier Linacre House, Jordan Hill, Oxford, 2nd edition, 2008.

[10] Ph. Kruchten, P. Lago, and H. Van Vliet. *Quality of Software Architectures*, volume 4214, chapter Building Up and Reasoning About Architectural Knowledge, pages 43 – 58. Springer Berlin Heidelberg, 2006.

[11] N. Leeming. Business process management implementation. *The Journal of Enterprise Architecture*, 1:69–91, 2005.

[12] A. Majchrzak, L.P. Cooper, and O.E. Neece. Knowledge reuse for innovation. *Management Science*, 50(2):174–188, 2004.

[13] M. L. Markus. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Joutnal of Management of Information Systems*, 18(1):57–93, 2001.

[14] K. McCormack and W. Johnson. *Business Process Orientation: Gaining the E-Business Competitive Advantage*. St Lucie Press, Delray Beach, FL, 2000.

[15] H. Mintzberg. Patterns in strategy formation. *Management Science*, 24(9):934–948, 1978.

[16] Bumiller J. Mutschler, B. Towards an evaluation framework for business process integration and management. In *2nd International Workshop on Interoperability Research for Networked Enterprise Applications and Software, EDOC*, Enschede, Netherlands, 2005. IEEE.

[17] Reijers H.A. van der Aast W.M.P. Netjes, M. Supporting the bpm life-cycle with filenet. *http://www.is.win.tue.nl/ wvdaalst/publications/p328.pdf*, 2006.

[18] E. Olding. Starting up the business process comptency center. *http://www.gartner.com*, 2007.

[19] E. Olding and J Hill. Role definition and organizational structure: Business process improvement. *Gartner*, 2008.

[20] OMG. Organization structure metamodel (osm). *http://www.omg.org/techprocess/meetings/schedule*, 2006.

[21] J. Recker. Opportunities and constraints: the current struggle with bpmn. *Business Process Management Journal*, 16(1):181–201, 2010.

[22] H.A. Reijers. Implementing bpm systems: The role of process orientation. *BPM Journal*, 12(4):389–409, 2006.

[23] M. Rosemann. Application of a holistic model for determining bpm maturity. *BPTrends*, Feb.:http://bpm–training.com/wp–content/uploads/2010/04/applicationholistic.pdf, 2005.

[24] J.H. Snabe, A. Rosenberg, C. Mžller, and M. Scavillo. *Business process management: the SAP roadmap.* Galileo Press, 2008.

[25] P. Trkman. The critical success factors of business process management. *International Journal of Information Management*, 30:125–134, 2010.

[26] W. van der Aalst, M. ter Hofstede, A., and M. Weske. Business process management: A survey, 2003.

[27] M. Weske. *Business Process Management: Concepts, Languages, Architectures*, volume ISBN-10: 3540735216. Springer, 2007.

[28] O. Zimmermann, Th. Geschwind, J. Kuester, F. Leymann, and N. Schuster. Reusable architectural decision models for enterprise application development. In S. Overhage, C. Szyperski, R. Reussner, and J. Stafford, editors, *Software Architectures, Components, and Applications*, volume 4880 of *Lecture Notes in Computer Science*, pages 15–32. Springer, 2007.

[29] F. Zucchi and J. S. Edwards. Human resource management aspects of business process reengineering: a survey. *Business Process Management Journal*, 5(4):325 – 344, 1999.

[30] Rosemann M. zur Muehlen, M. Multi-paradigm process management. In M. Kirikova J. Grundspenkis, editor, *Proceedings of CAiSE'04 Workshops - 5th Workshop on Business Process Modeling, Development and Support (BPMDS 2004)*, pages 169–175, Riga, Latvia, 2004.

 ISBN: 978-1-61208-160-1

Design Engineering Practices in Indian Manufacturing Firms: An Empirical Study

Santanu Roy
Institute of Management Technology (IMT)
Hapur Road,
Raj Nagar,
Ghaziabad 201 001
India
sroy@imt.edu; rsan58@yahoo.co.uk

Parthasarathi Banerjee
National Institute of Science, Technology and Development Studies (NISTADS)
K.S. Krishnan Marg, Pusa,
New Delhi 110 012
India
psb_nist@yahoo.com

Abstract **- The design process is a critical component in competitive product development and in the industrial innovation process. The work reported in the paper tries to map the different patterns of design engineering practices as industrial innovation indicators as they occur in firms located within the National Capital Region (NCR) of India representing New Delhi and its surrounding regions and highlights the role of developing suitable indicators to tap specific design engineering practices and the network linkages. The results indicate that majority of firms possessing a separate design department exhibit a better appreciation of what constitutes a successful innovation and follow it up by formulating design engineering agreements with a networked partner simultaneous to marketing, service or R&D arrangements, and also that firms with more open cooperation with the outside research environments almost always have been more technically successful in designing new products.**

Keywords - design engineering; national capital region; India; indicators; industrial innovation.

1. INTRODUCTION

National governments in many countries launch specific schemes in order to promote innovative activities in the manufacturing sectors like providing fiscal incentives for research and development (R&D) and quality control [1, 2, 3]. This is especially true for developing countries like India. Such schemes are often monitored through indicators devised for this purpose.

The significance of the term design, its definition, its conceptual construct, and also the critical role it plays in innovation policy framework, are increasingly being recognized [4, 6, 8, 9, 10]. A very concise definition of design describes it as a creative process by which product innovations and ideas are reduced to an economically viable arrangement, this arrangement being set down on paper as a proper schedule. The concept of economic viability can be expanded to include (i) the economic objective; (ii) a product to meet that objective; (iii) how effectively the product meets the economic objective; (iv) how well the product work; (v) how the product can be made; (vi) hence, how much the product will cost to make; and (vii) what the product will cost to maintain. When a firm moves into new product areas and technologies, and as the firm's competitive context that engineering design has a strong relationship with the management practices being adopted for successful innovation, and plays a significant part in improving the competitiveness of products or firms [4, 5]. Li and Boyle [6] presents review of papers published in the Journal of Engineering Design from 2007-2008 probing the perspectives, challenges and recent advances in engineering design.

The aforementioned discussion clearly brings into focus the imperative and the need to bring design, consciously, within the framework of an innovation policy. A manufacturing firm that seeks to compete effectively in the market needs to formulate and implement innovation-based strategies where design forms an integral part, for instance, in developing new products and in meeting customer satisfaction, but empirical or theoretical research on design engineering practices as indicators of industrial innovation have been few. The work reported in the present paper aims to fill this gap. The study attempts to map the different patterns of design engineering as are being practiced in the firms located within the National Capital Region (NCR) of India, meaning New Delhi and its surrounding areas. After a thorough review of the literature in the next section, methods employed and the data sources have been explained in the following section. The last two sections deal with the results of the analysis and the broad conclusions of the study, respectively.

2. DESIGN IN THE CONTEXT OF INDUSTRIAL INNOVATION

When a firm moves into new product areas and technologies, and as the firm's competitive context becomes less predictable and more complex, communication between the firm and the world outside will increase. Kalogerakis, Luthje, and Herstatt [7] report a work where in-depth interviews were carried out with project leaders of 18 design and engineering consulting firms located in Germany and Scandinavia in order to probe the links between design innovations and analogies. Other studies have highlighted that engineering design has a strong relationship with the management practices being adopted for successful innovation, and plays a significant part in improving the competitiveness of products, firms or even national economies [8, 9, 10]. In this regard, it

 ISBN: 978-1-61208-160-1

could be pertinent to mention that a special issue of the International Journal of Production Research on data mining applications in engineering design, manufacturing and logistics was brought out in the year 2006 [11]. There is also the issue of outsourcing the engineering design process by a firm. The limits of design and engineering outsourcing in product development, and the sources of these limits have been a subject of debate [12]. About 100 research papers and 30 commercial systems/international standards launched have been reviewed in terms of underlying algorithms, mechanisms and system architectures with a view to focus on research being carried out to develop methodologies and technologies to support geographically dispersed teams to organise collaborative design based on the quickly evolving information technologies [13].

The role of design engineering becomes very critical in cases where the project requires too advanced a technology and/or the transition to manufacturing is complex and requires very large scale-up. This is typically true for technology transfer agreements from the national R&D laboratories to the industries. The products and the processes are found to be most satisfactory in the bench-scale on lab-scale work but in the commercial scale operations, these often fail to live up to the expectations and the innovative efforts might lead to a failure [14, 15, 16, 17]. The issues raised, therefore, focus on identifying appropriate customers and working closely with them [18].

Other studies have highlighted the significance of design in the innovation process and the criticality of recognizing the same. By focusing on the environmental impacts of the electronics industry, the perspectives for design for environment have been debated [19]. In a separate study of 203 new products - both winners and losers that were launched into the market place - three hypothesized factors, all of them directly or indirectly related to design were found to be most significantly related to design new product success [20]. The factors were (i) product advantage (ii) proficiency of predevelopment activities, and (iii) protocol, including product concept, specifications and requirements. Among recent works in the Indian context, a case study was carried out in an Indian manufacturing organization for probing the role of computer aided design and engineering as enablers of agile manufacturing [21]. In the same vein, engineering firms in Thailand have been probed to propose a decision support methodology for the development and application of product eco-design, with special reference to their use in these firms [22].

Information about the user's needs helps to clarify the firm's activity. A firm that has achieved a through understanding of the needs of the buyer can use the knowledge to guide innovative effort. A firm that faces a technology adoption decision engage in an extension effort reduce uncertainty associated with that decision [23], and information about the buyer is an important part of this exercise. In situations where the clarity of innovation objectives is higher, the firm is expected to be more willing to engage in innovation. The information-seeking networks of marketing managers are closely tied to sources that clarify the customer's needs. When a firm moves into new product areas and technologies, and as the firm’s competitive context becomes less predicable and more complex, communications between the firm and the world outside will increase [24].

It has been observed that one of the most important factors affecting the competitiveness of American manufacturing sectors is the long-standing neglect of engineering design [25]. The authors assert the following: (i) in terms of long-range strategy, the factors of cost, quality and time-to-market are design problems more than manufacturing problems; (ii) market loss by US companies is due to design deficiencies more than manufacturing deficiencies; (iii) manufacturing process themselves are designed; (iv) many technical problems commonly associated with manufacturing process are traceable to design problems; and (v) opportunities to surpass foreign competitors are best found in engineering design. Spitas [26] has reported a work carried out through a questionnaire-based survey of design engineers to evaluate the industry’s perception and use of systematic design paradigms.

As already mentioned, a particular new product failure might act as the seed for the germination of successful redesigned product. The design process is an important stage in new product development. Based on graph theory and the weighting concept, a quantified design structure matrix (a systematic planning method of optimizing design priorities and product architecture for managing product variety from an informational structure perspective) has been presented in the literature [27]. In another study, the relationship between different retail channel structures and channel strategies (for instance, an exclusive channel strategy) and the engineering design of a new product, conditional on consumer preference distributions and competitor product attributes have been examined [28]. Even other researchers have emphasized that by and large designs are modifications from previous products and lessons learned from earlier designs can be beneficial when developing new products [29]. Based on the example of a new generation of diesel engine design, the authors have shown how the ability to predict change propagation can guide designers through conceptual design allowing them to analyse design alternatives and foresee potential problems arising from the product architecture.

 ISBN: 978-1-61208-160-1

3. METHODS/DATA

For the study, a stratified sample out of a population of industrial concerns operating within the National Capital Region (NCR) of India has been considered. Data were collected from 53 firms operating within the NCR of varying annual turnover, size and belonging to automobiles, engineering, chemicals/pharmaceuticals/textiles, and electrical/electronics sectors. A questionnaire, formulated on our preliminary understanding of the innovation process, was later modified with inputs from a pilot survey to be used in the later part of field investigations. The responses were subject to detailed analysis in terms of dimensions like design objectives, sources of information, network partners in the design processes, etc.

The analysis has been carried out in various stages to bring out the critical importance of design engineering in industrial innovation, and the factors related to the success of innovative efforts. Category 1 (numbering 23) firms refer to those firms that have a separate design department while Category 2 (numbering 30) firms have no separate design department.

4. RESULTS

Table 1 presents the total value of machinery as a percentage of gross turnover of the firms. Set against the importance of establishing and running a separate design engineering department in a firm, it is observed from the Table that this percentage value is 15.58% for firms with a separate design department that is substantially lower as compared to 33.59% for firms without a separate design department. On a closer examination, it is found that this difference is most conspicuous for small (up to INR 20 million turnover) and very large firms (a turnover of higher than INR 200 million) where 45 INR=1 USD.

Further analysis has been carried out separately for Category 1 and Category 2 firms.

Table 2 presents the extent of database maintenance by different industrial sectors. It is observed from this Table that except for the chemicals and pharmaceutical sector industries, all others maintain engineering databases to a reasonable extent.

TABLE 1. TOTAL VALUE OF MACHINERY AS PERCENTAGE OF GROSS TURNOVER

Turnover (in million INR)	**Category 1**	**Category 2**
Up to 20	14.05	34.81
20-99	19.26	22.16
100-200	16.12	37.42
More than 200	12.88	39.97
Average	15.58	33.59

* 45 INR = 1USD

On the contrary, formulation databases are not maintained to any significant extent by any industrial sector. Although all the sectors maintain manufacturing databases, the percentage of firms maintaining such a database is indeed very high for the automobile sector (77.36%).

This sector exhibits a similar response regarding maintenance of inventory databases that otherwise are also maintained by engineering sector industries to a fair extent but not by chemicals/pharmaceuticals and electrical/electronics sector industries. Significantly, marketing databases are not maintained by automobiles and engineering sectors and these are maintained only to a limited extent by the other sectors.

Table 3 presents the design objectives of the firms (first choice objectives). It is found that for firms possessing a design department (category 1 firms), meeting unique demands of the customer is the predominant design objective (79.2%) whereas reliability (13.9%) is also important. For firms without a separate design department (category 2 firms), the predominant first choice design objective is meeting unique customer demands whereas minimum consumption of materials and resources, surpassing the features of competitors' products, ergonomics and ease of operation, optimality, and reliability are also relevant.

Table 4 presents the second choice design objectives. For the first category of firms, ensuring minimum consumption of materials and resources is the most important design objective. Other prominent design objectives for this category include ergonomics and ease of operation (21.0%), optimality (14.2%), and surpassing the features of competitors' products (14.2%). For category 2 firms, the prominent ones include ease of manufacturing, surpassing the features of competitors' products (that is, in fact common to both categories of firms), and minimum consumption of materials and resources but the most significant one is reliability (45%) of the designed product.

A. Design and industrial innovation: Significance of networks

Table 5 through Table 8 help illustrate the significance of networks in technological innovation. Aspects of design engineering as industrial innovation indicators do not act in isolation but linked with other relevant actors.

Table 5 presents the first choice sources of information for the design function. For category 1
firms, the marketing team is the most important source of information (60.7%) followed by dealers (32.2%). There are the only two prominent sources for this category of firms.

 ISBN: 978-1-61208-160-1

TABLE 2. DATABASE MAINTENANCE

Firm type	Percentage of firms (Engineering)	Percentage of firms (Formulation)	Percentage of firms (Manufacturing)	Percentage of firms (Inventory)	Percentage of firms (marketing)
Automobiles	66.03	22.64	77.36	77.36	33.96
Engineering	54.71	27.36	54.71	45.28	27.36
Chemicals/pharmaceuticals/ textiles	28.30	14.16	42.54	28.30	42.54
Electrical/electronics	66.03	0.0	55.56	22.64	44.54

TABLE 3. DESIGN OBJECTIVES (FIRST CHOICE)

Firm Category	Meeting unique customer demands	Minimum consumption of materials/ Resources	Surpass the features of competitors' products	Ease of manufacture	Ergonomics and ease of operation.	Opti-mality	Reliability
Cat.1	79.2	0	0	0	6.9	0	13.9
Cat. 2	47	13	13	0	13	6.5	6.5

TABLE 4. DESIGN OBJECTIVES (SECOND CHOICE)

Firm Category	Meeting unique customer demands	Minimum consumption of materials/ Resources	Surpass the features of competitors' products	Ease of manufacture	Ergonomics and ease of operation.	Opti-mality	Reliability
Cat.1	7.4	28.5	14.2	7.4	21.0	14.2	7.4
Cat. 2	0	11	22	22	0	0	45

TABLE 5. SOURCES OF INFORMATION (FIRST CHOICE)

Firm Category	Mkt. Team	Dealers	Mkt. survey	Vendors	Service team	Joint venture/network partner
Category 1	60.7	32.2	7.1	0	0	0
Category 2	37.7	13.2	5.7	18. 8	13.2	13.2

TABLE 6. SOURCES OF INFORMATION (SECOND CHOICE)

Firm Category	Mkt. Team	Dealers	Mkt. survey	Vendors	Service team	Joint venture/network partner
Category 1	30.6	23.8	14.3	9.5	14.3	9.5
Category 2	13.6	27.3	13.6	9. 1	22.7	13.6

TABLE 7. WHETHER DESIGNING AGREEMENT/ARRANGEMENT WITH A NETWORKED PARTNER IS SIMULTANEOUS TO OTHER PARALLEL AGREEMENTS

Firm Category	Marketing arrangement	Service/training arrangement	R&D arrangement	Design engineering arrangement
Category 1	36.4	36.4	42.2	42.2
Category 2	5. 0	15. 0	15. 0	15. 0

TABLE 8. WHO GENERALLY PROVIDES THE SPECIFICATIONS AND ENGINEERS THE PROTOTYPES

Firm Category	Joint venture partner/ technical collaborators	Own design team/ other in-house team	Business associate/ partner	Engineering consultant	R&D consul-tant	Software consul-tant	Outsider
Cat.1	8.3	75	25	25	16.6	16.6	32.3
Cat. 2	19	38.1	0. 0	4.8	4.8	0. 0	19

 ISBN: 978-1-61208-160-1

For category 2 firms, this distribution is more widespread with marketing team (37.7%), dealers (13.2%), vendors/suppliers (18.8%), the service team (13.2%), and joint venture/network partner (13.2%) as important sources of information.

Table 6 presents the second choice sources of information for the design function. Regarding this second choice, the important ones both for category 1and 2 firms include the marketing team, the dealers, market surveys and the service team. The only additional sources of information of category 2 firms are the joint venture/network partners.

Table 7 displays what other agreements are simultaneous to a designing agreement or arrangement among firms and their network partners. The findings re-emphasize of a network approach for successful industrial innovation. For firms with a separate design department (category 1), the design agreements are often simultaneous to a marketing arrangement (36.4%), servicing/training arrangement (36.4%), R&D arrangement (42.2%), and design engineering arrangement (42.2%). For category 2 firms, all these arrangements (except marketing arrangement) are at times entered into but much less frequently (15%).

Illustrating this network approach further, Table 8 points out that among all the network partners of a firm, except for joint venture partner/technical collaborators, all others take part in the design of the prototypes for category 1 firms. For such firms, the in-house design departments are most of the time doing this job (75.0%), though there are other categories too like the business associates/partners (25.0%), the engineering consultants (25.0%), R&D or software consultants (16.6%), and even outsiders (32.3%). For category 2 firms as well, the prototypes are mostly designed in-house (38.1%) but this percentage is much lower than that for firms which do possess a separate design department. Others who design the prototypes include the joint venture partners/technological collaborators (19.0%), and outsiders (19.0%).

5. CONCLUSION

The study thus highlights the critical aspects of the design process as industrial innovation indicators. These results can be related to some other studies on different aspects of the design process. Reference could be made to the results of a study conducted in a medium-sized aerospace company according to which the percentage of time spent by engineers in different activities had a high component of documentation (28.5%), solving thinking (28.0%), support and consulting (17.1%), and information gathering (13.7%) [27]. In yet another study, it was shown that competitiveness of a manufactured product can be improved by (i) good product design; (ii) product innovation, and (iii) production process improvements [28]. Their study results have also shown that product design could affect both price competition through design for economic manufacture and low life-cycle costs, and non-price competition, either through the technical design of the product itself to improve performance, appearance, quality, etc., or by taking into account associated service-related non-price factors.

The results of our study, similarly, are in line with the findings of another empirical study carried out that has shown that two main strategic dimensions are related to success in technically designing new products [29]. The first is the R&D orientation of the firm and the second is the technology use. The data shows that regardless of industry, firms with more open cooperation with the outside research environment, that is more external R&D strategies, almost always have been more technically successful in designing new products. Network positioning, gaining access to external information and assistance by having wide and flexible contacts with the external environment was, therefore, one of the most strategic variables in their analysis. The results of our study corroborate the above and establish the significance of the network approach in formulating a design engineering policy for a firm that promotes industrial innovation.

The results of the study indicate that design engineering plays a critical role in fostering innovations in industrial firms. It highlights the role of developing suitable indicators to tap specific design engineering practices, the network linkages, and what all factors need to be looked into while initiating policy measures to promote industrial innovation in the manufacturing sectors. It has been observed, for instance, that majority of firms possessing a separate DD exhibit a better appreciation of what constitutes a successful innovation and follow it up by formulating design engineering agreements with a networked partner simultaneous to marketing, service or R&D arrangements. The results have implications for formulating policies that seek to promote industrial innovation and the incentive schemes that go along with it. Industrial innovation strategies in force at the national or local levels have largely overlooked the importance of design engineering in fostering industrial innovation and the process intricacies inherent in it. The present study aims to do just that. There is, therefore, a need for a

 ISBN: 978-1-61208-160-1

more extensive empirical research on a larger sample size and spread over a larger geographical spread to further consolidate the insights gained from the present study.

6. REFERENCES

[1] P. Biegelbauer and S. Borras, Innovation Policies of Europe and US: The new Agenda, England: Ashgata Publishing, 2003.

[2] S. Borras, Innovation Policy of the EU: From Govt to Governance, UK: Edward Elgar, 2003.

[3] F. Malerba and S. Brusoni, Perspective on Innovation, London: Cambridge University Press, 2007.

[4] R. Roy and S. Potter, "The commercial impacts of investments in design," Design Studies, vol. 14(2), pp. 171-193, 1993.

[5] R. Roy, "Can the benefits of good design be quantifies?," Design Management Journal, vol. 5(2), pp. 9-17, 1994.

[6] S. Li and I.M. Boyle, "Engineering design: Perspectives, challenges, and recent advances," Journal of Engineering Design, vol. 20(1), pp. 7-19, 2009.

[7] K. Kalogerakis, C. Luthje, and C. Herstatt, "Developing innovations based on analogies: Experience from design and engineering consultants," Journal of Product Innovation Management, vol. 27, pp. 418-436, 2010.

]8] C. Freeman, The Economics of Hope: Essays on Technical Change, Economic Growth and the Environment, London: Frances Pinter, 1992.

[9] J.M. Utterback, Mastering the Dynamics of Innovation, Boston, MA: Harvard Business School Press, 1994.

[10] T. Peters, "Pundit of passion", Design, vol. Summer, pp. 18-19, 1995.

[11] J.C.X. Feng and A. Kusiak, A., "Data mining applications in engineering design, manufacturing and logistics: Guest editorial," International Journal of Production Research, vol. 44(14-15), pp. 2689-2694, 2006.

[12] F. Zirpoli and M.C. Becker, "The limits of design and engineering outsourcing: Performance integration and unfulfilled promises of modularity," R&D Management, vol. 41(1), pp. 21-43, 2011.

[13] W.D. Li. and Z.M. Qiu, "State-of-art technologies and methodologies for collaborative product development systems," International Journal of Production Research, vol. 44(13), pp. 2525-2559, 2006.

[14] S. Roy, "Technology transfer from R&D centres in India and development of industrial clusters," in Enterprise Support Systems: An International Perspective, M.J. Manimala, J. Mitra and V. Singh, Eds., ISBN: 978-81-7829-927-3, New Delhi: Sage, 2009, pp. 217-227.

[15] S. Roy and P. Banerjee, "Developing regional clusters in India: The role of national laboratories," International Journal of Technology Management and Sustainable Development, vol. 6(3), pp. 193-210, 2007.

[16] S. Roy, "Networking as a strategy for technology transfer and commercialization from R&D laboratories," Industry and Higher Education, vol. 20(2), pp. 123-133, 2006.

[17] S. Roy and P.K.J. Mohapatra, "Regional specialisation for technological innovation in R&D laboratories: A strategic perspective," Artificial Intelligence and Society, vol. 16, pp. 100-111, 2002.

[18] E. von Hippel, "Lead users: A source of novel product concepts," Management Science, vol. 32(7), pp. 791-805, 1986.

[19] C. Boks and A. Stevels, "Essential perspectives for design for environment: Experiences from the electronics industry," International Journal of Production Research, vol. 45(18-19), pp. 4021-4039, 2007.

[20] R.G. Cooper and E.J. Kleinschmidt, "New products: What separates winners from losers?," in Managing Innovation, J. Henry and D. Walker, Eds., Sage: London, 1991.

[21] S. Vinodh and D. Kuttalingam, "Computer aided design and engineering as enablers of agile manufacturing: A case study in an Indian manufacturing organization," Journal of Manufacturing Technology Management, vol. 22(3), pp. 405-418, 2011.

[22] P. Boonkanit and A. Kengpol, "The development and application of a decision support methodology for product eco-design: A study of engineering firms in Thailand," International Journal of Management, vol. 27(1), pp. 185-199, 2010.

[23] K.F. McCardle, "Information acquisition and the adoption of new technology," Management Science, vol. 31(11), pp. 1372-1389, 1985.

[24] J.W. Brown and J.M. Utterback, "Uncertainty and technical communication patterns," Management Science, vol. 31(3), pp. 301-311, 1985.

[25] J.R. Dixon and M.R. Duffey, "The neglect of engineering design," California Management Review, vol. 32(2), pp. 9-23, 1990.

[26] C. Spitas, "Analysis of systematic engineering design paradigms in industrial practice: A survey," Journal of Engineering Design, vol. 22(6), pp. 427-445, 2011.

[27] D. Luh, Y. Ko, and C. Ma, "A structural matrix-based modelling for designing product variety," Journal of Engineering Design, vol. 22(1), pp. 1-29, 2011.

[28] N. Williams, P.K. Kannan, and S. Azarm, "Retail channel structure impact on strategic engineering product design," Management Science, vol. 57(5), pp. 897-914, 2011.

[29] R. Keller, C.M. Eckert, and P.J. Clarkson, "Using an engineering change methodology to support conceptual design," Journal of Engineering Design, vol. 20(6), pp. 571-587, 2009.

[30] R.A. Crabtree, M.S. Fox, and N.K. Baid, "Case studies of coordination activities and problems in collaborative design," Research in Engineering Design, vol. 9**,** pp. 70-84, 1997.

[31] R. Roy and J.C. Riedel, "The role of design and innovation in competitive product development," Second European Academy of Design Conference: Contextual Design – Design in Contexts, Stockholm, Sweden, 23-25 April, 1997.

[32] H. Nystrom, "Company strategies for designing and marketing new products in the electrotechnical industry," in Design and Innovation: Policy and Management, R. Longdon and R. Rothwell, Eds., London: Frances Pinter, 1985, pp. 18-26.

 ISBN: 978-1-61208-160-1

ValidKI: A Method for Designing Key Indicators to Monitor the Fulfillment of Business Objectives

Olav Skjelkvåle Ligaarden*†, Atle Refsdal*, and Ketil Stølen*†
* *Department for Networked Systems and Services, SINTEF ICT, PO Box 124 Blindern, N-0314 Oslo, Norway*
E-mail: {olav.ligaarden, atle.refsdal, ketil.stolen}@sintef.no
† *Department of Informatics, University of Oslo, PO Box 1080 Blindern, N-0316 Oslo, Norway*

***Abstract*—In this paper we present our method ValidKI for designing key indicators to monitor the fulfillment of business objectives. A set of key indicators is valid with respect to a business objective if it can be used to measure the degree to which the business or relevant part thereof complies with the business objective. ValidKI consists of three main steps each of which is divided into sub-steps. We demonstrate the method on an example case focusing on the use of electronic patient records in a hospital environment.**

***Keywords*-key indicator, business objective, electronic patient record**

I. INTRODUCTION

Today's companies benefit greatly from ICT-supported business processes, as well as business intelligence and business process intelligence applications monitoring and analyzing different aspects of a business and its processes. The output from these applications may be key indicators which summarize large amounts of data into single numbers. Key indicators can be used to evaluate how successful a company is with respect to specific business objectives. For this to be possible it is important that the key indicators are valid. A set of key indicators is valid with respect to a business objective if it can be used to measure the degree to which the business or relevant part thereof complies with the business objective. Valid key indicators facilitate decision making, while invalid key indicators may lead to bad business decisions, which again may greatly harm the company.

In today's business environment, companies cooperate across company borders. Such co-operations often result in sharing or outsourcing of ICT-supported business processes. One example is the interconnected electronic patient record (EPR) infrastructure. The common goal for this infrastructure is the exchange of EPRs facilitating the treatment of the same patient at more than one hospital. In such an infrastructure, it is important to monitor the use of EPRs in order to detect and avoid misuse. This may be achieved through the use of key indicators. It may be challenging to identify and compute good key indicators that are valid. Furthermore, in an infrastructure or system stretching across many companies we often have different degrees of visibility into how the cooperating parties perform their part of the business relationship, making the calculation of key indicators particularly hard.

In this paper we present a new method *ValidKI* (Valid Key Indicators) for designing key indicators to monitor the fulfillment of business objectives. We demonstrate ValidKI by applying it on an example case targeting the use of EPRs. We have developed ValidKI with the aim of fulfilling the following characteristics:

- **Business focus:** The method should facilitate the design and assessment of key indicators for the purpose of measuring the fulfillment of business objectives.
- **Efficiency:** The method should be time and resource efficient.
- **Generality:** The method should be able to support design of key indicators for systems shared between many companies or organizations.
- **Heterogeneity:** The method should not place restrictions on how key indicators are designed.

To the best of our knowledge, there exists no other method with sole focus on design of valid key indicators to monitor the fulfillment of business objectives.

The rest of the paper is structured as follows: in Section II we introduce our basic terminology and definitions. In Section III we give an overview of ValidKI and its three main steps. In Sections IV, V, and VI we demonstrate our three-step method on an example case addressing the use of EPRs in a hospital environment. In Section VII we present related work, while in Section VIII we conclude by characterizing our contribution and discussing the suitability of our method.

II. BASIC TERMINOLOGY AND DEFINITIONS

Merriam-Webster defines an "indicator" as *"one that indicates"* [1], while it defines "indicates" as *"to be a sign, symptom, or index of"* [2]. The weather forecast is a typical indicator since it gives an indication of what the weather will be like the next day.

Many companies profit considerably from the use of indicators [3] resulting from business process intelligence applications that monitor and analyze different aspects of a business and its processes. Indicators can be used to measure to what degree a company fulfills its business objectives and

 ISBN: 978-1-61208-160-1

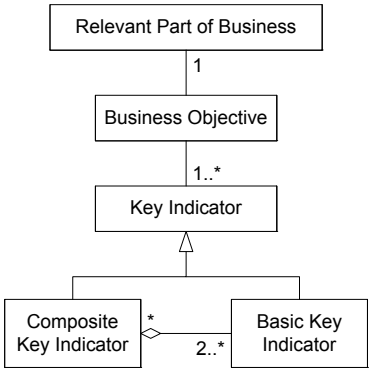

Figure 1. The artefacts addressed by ValidKI

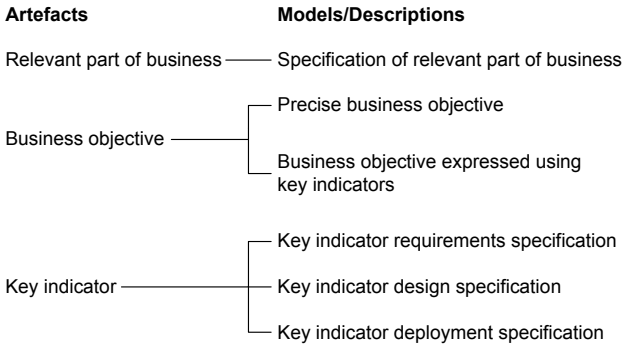

Figure 2. The models/descriptions developed by ValidKI

we then speak of key indicators. Some business objectives may focus on business performance, while others may focus on risk or compliance with laws and regulations.

A. The artefacts addressed by ValidKI

The UML [4] class diagram in Fig. 1 relates the main artefacts addressed by ValidKI. The associations between the different concepts have cardinalities that specify how many instances of one concept that may be associated to an instance of the other concept. The hollow diamond specifies aggregation.

As characterized by the diagram, a key indicator is either basic or composite. By basic key indicator we mean a measure such as the number of times a specific event generated by the ICT infrastructure has been observed within a given time interval, the average time between each generation of a specific event, the load on the network at a particular point in time, or similar. A composite key indicator is the aggregation of two or more basic key indicators. One or more key indicators are used to measure to what extent a business objective is fulfilled with respect to a relevant part of the business.

B. The models/descriptions developed by ValidKI

As illustrated by Fig. 2, performing the steps of ValidKI results in six different models/descriptions each of which describes one of the artefacts of Fig. 1 from a certain perspective.

A specification, at a suitable level of abstraction, documents the relevant part of the business in question.

Business objectives are typically expressed at an enterprise level and in such a way that they can easily be understood by for example shareholders, board members, partners, etc. It is therefore often not completely clear what it means to fulfill them. This motivates the need to capture each business objective more precisely. The degree of fulfillment of a precise business objective is measured by a set of key indicators. To measure its degree of fulfillment there is a need to express each business objective in terms of key indicators.

For each key indicator we distinguish between three specifications; the key indicator requirements specification, the key indicator design specification, and the key indicator deployment specification. The first captures the expectations to the key indicator, the second defines how the indicator is calculated, while the third documents how the calculation is embedded in the business or relevant part thereof that ValidKI is used to help monitor. In other words the deployment specification describes how the data on which the key indicator calculation is based is extracted and transmitted within the business in question. We often refer to the pair of the design specification and the deployment specification as the key indicator's realization.

C. External and internal validity

We distinguish between external and internal validity. External validity may be understood as a relation between a business objective and a set of key indicators.

Definition 1. ***External validity*** *A set of key indicators is externally valid with respect to a business objective if it can be used to measure the degree to which the business objective is fulfilled.*

Internal validity may be understood as a relation between the requirements specification of a key indicator and its realization as captured by its design and deployment specifications.

Definition 2. ***Internal validity*** *A key indicator is internally valid if its realization as captured by its design and deployment specifications fulfills its requirements specification.*

If each element of a set of key indicators is internally valid we may evaluate its external validity by considering only the requirements specifications of its elements.

III. Overview of ValidKI

Fig. 3 provides an overview of the ValidKI method. It takes as input a business objective and delivers a set of key indicators and a report arguing its external validity with

 ISBN: 978-1-61208-160-1

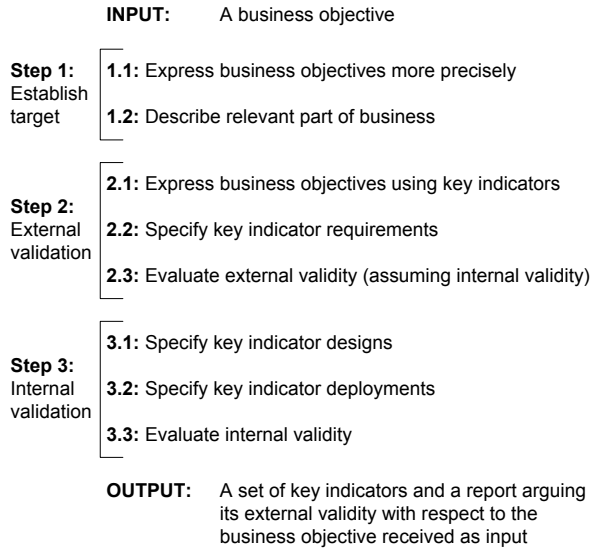

Figure 3. Overview of ValidKI

respect to the business objective received as input[1]. In the following we offer additional explanations for each of the three main steps.

A. Establish target

The first main step of ValidKI is all about understanding the target, i.e. understanding exactly what the business objective means and acquiring the necessary understanding of the relevant part of business for which the business objective has been formulated. In the first sub-step we help the client to characterize the business objective in a more precise manner leading to a precise business objective, while in the second sub-step we specify the relevant part of the business.

B. External validation

The second main step of ValidKI is concerned with establishing a set of key indicators that is externally valid with respect to the business objective considering only their requirements specifications.

In order to argue external validity we reformulate the precise business objective in terms of key indicators. Furthermore, we specify our requirements to each key indicator referred to in the reformulated precise business objective. These two sub-steps are typically conducted in parallel. Based on this we evaluate external validity assuming internal validity of each key indicator. If we are not able to establish external validity the reformulated business objective and/or requirements specifications must be changed.

[1]When using ValidKI in practice we will typically develop key indicators for a set of business objectives, and not just one which we without loss of generality restrict our attention to here.

C. Internal validation

In the third main step we do an internal validation of each key indicator of the externally valid key indicator set. In the first two sub-steps we specify the design and deployment of each key indicator. Then, we evaluate whether this realization fulfills its requirements specification. If this is the case we have established internal validity and thereby external validity; if not we iterate.

IV. ESTABLISH TARGET

In the following we assume that we have been hired to help the public hospital Client H design key indicators to monitor their compliance with Article 8 in the European Convention on Human Rights [5]. The article states the following:

Article 8 – Right to respect for private and family life

1) Everyone has the right to respect for his private and family life, his home and his correspondence.
2) There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Client H needs to comply with Article 8 since it is a public authority. The consequence for Client H of not complying with Article 8 may be economic loss and damaged reputation. One example [6] of violation of Article 8 is from Finland. A Finnish woman was first treated for HIV at a hospital, before she later started working there as a nurse. While working there she suspected that her co-workers had unlawfully gained access to her medical data. She brought the case to the European Court of Human Rights in Strasbourg which unanimously held that the district health authority, responsible for the hospital, had violated Article 8 by not protecting the medical data of the woman properly. The district health authority was held liable to pay damages to the woman. Client H has therefore established the following business objective:

Business objective BO-A8: Client H complies with Article 8 in the European Convention on Human Rights.

Client H wants to make use of key indicators to monitor the degree of fulfillment of BO-A8, and now they have hired us to use ValidKI for designing them. In the rest of this section we conduct Step 1 of ValidKI on behalf of Client H with respect to BO-A8.

 ISBN: 978-1-61208-160-1

A. Express business objectives more precisely (Step 1.1 of ValidKI)

Article 8 states under which circumstances a public authority can interfere with someone's right to privacy. One of these circumstances is *"for the protection of health"*, which is what Client H wants us to focus on. In the context of Client H this means to provide medical assistance to patients. The ones who provide this assistance are the health-care professionals of Client H.

The medical history of a patient is regarded as both sensitive and private. At Client H, the medical history of a patient is stored in an electronic patient record (EPR). An EPR is *"an electronically managed and stored collection or collocation of recorded/registered information on a patient in connection with medical assistance"* [7]. The main purpose of an EPR is to communicate information between health-care professionals that provide medical care to a patient. To protect the privacy of its patients, Client H restricts the use of EPRs. In order to comply with Article 8, Client H allows a health-care professional to interfere with the privacy of a patient only when providing medical assistance to this patient. Hence, the dealing with the EPRs within the realms of Client H is essential.

For Client H it is important that every access to information in an EPR is in accordance with Article 8. A health-care professional can only access a patient's EPR if he/she provides medical assistance to that patient, and he/she can only access information that is necessary for the medical assistance provided to the patient. The information accessed can not be used for any other purpose than providing medical assistance to patients. Accesses to information in EPRs not needed for providing medical assistance would not be in accordance with Article 8. Also, employees that are not health-care professionals and that work within the jurisdiction of Client H are not allowed to access EPRs. Based on the constraints provided by Client H, we decide to express BO-A8 more precisely as follows:

Precise business objective PBO-A8: $C_1 \wedge C_2 \wedge C_3$

- **Constraint** C_1**:** Health-care professionals acting on behalf of Client H access:
  - a patient's EPR only when providing medical assistance to that patient
  - only the information in a patient's EPR that is necessary for providing medical assistance to that patient
- **Constraint** C_2**:** Health-care professionals acting on behalf of Client H do not use the information obtained from a patient's EPR for any other purpose than providing medical assistance to that patient.
- **Constraint** C_3**:** Employees that are not health-care professionals and that work within the jurisdiction of Client H do not access EPRs.

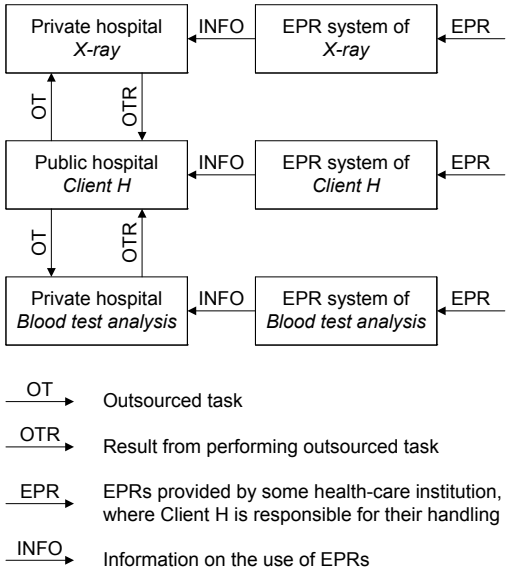

Figure 4. Specification of relevant part of business

As indicated by PBO-A8's definition, all three constraints must be fulfilled in order for PBO-A8 to be fulfilled.

B. Describe relevant part of business (Step 1.2 of ValidKI)

To design key indicators to monitor BO-A8 we need to understand the part of business that is to comply with BO-A8 and therefore is to be monitored. Client H has outsourced some of its medical services to two private hospitals. These two are referred to as *X-ray* and *Blood test analysis* in Fig. 4. The first hospital does all the X-ray work for Client H, while the second hospital does all the blood test analyses. Client H is not only responsible for its own handling of EPRs, but also the outsourcing partners' handling of EPRs, when they act on behalf of Client H. As shown in Fig. 4, Client H outsources tasks to the two private hospitals, and gets in return the results from performing these tasks. All three health-care institutions use some kind of EPR system for handling the EPRs. An EPR system is *"an electronic system with the necessary functionality to record, retrieve, present, communicate, edit, correct, and delete information in electronic patient records"* [7]. These systems use EPRs provided by several different health-care institutions. As shown in Fig. 4, these systems are only of interest when they handle EPRs where Client H is responsible for their handling. These systems will provide their institutions with information on the use of EPRs. This information can later be used in the monitoring. It should be noticed that the model in Fig. 4 only provides a small overview of the modeling that is performed.

V. External validation

In this step we establish a set of key indicators that is externally valid with respect to the business objective BO-

 ISBN: 978-1-61208-160-1

A8 by only considering their requirements specifications. In order to argue external validity we reformulate the precise business objective PBO-A8 in terms of key indicators. Due to lack of space, we only show how we reformulate constraint C_1.

A. Express business objectives using key indicators (Step 2.1 of ValidKI)

At the three health-care institutions, most of the medical tasks that a health-care professional conducts during a working day are known in advance. It is known which patients the professional will treat and what kind of information the professional will need access to in order to treat the different patients. When a health-care professional accesses information in a patient's EPR it is then possible to check whether the professional really needs this information. Client H and the two outsourcing partners have a list for each health-care professional documenting which patients the professional is treating and what kind of information the professional needs for this purpose. These lists are updated on a daily basis. Many of these updates are automatic. For instance, when Client H is assigned a new patient, then this patient is added to the lists of the health-care professionals that will be treating this patient.

The EPR systems classify an access to information in an EPR as *authorized* if the professional needs the information to do a planned task. Otherwise, the access is classified as *unauthorized*. If it is classified as unauthorized then it is possible to check in retrospect whether the access was necessary. In an emergency situation, for instance when a patient is having a heart attack, a health-care professional often needs access to information in an EPR that he/she was not supposed to access. By checking in retrospect whether unauthorized accesses were necessary it is possible to classify the unauthorized accesses into two groups; one for accesses that were necessary, and for those that were not. The first group is called *approved* unauthorized accesses, while the second group is called *not approved* unauthorized accesses. All accesses that are classified as not approved unauthorized accesses are considered as *illegal* accesses.

At Client H and the two outsourcing partners, health-care professionals use smart cards for accessing information in EPRs. If a card is lost or stolen, the owner must report it as missing, since missing cards may be used by other health-care professionals to access EPRs illegally[2]. When the card has been registered as missing it can no longer be used. When reporting it as missing, the last time the card owner used it before noticing that it was missing is recorded. All accesses to EPRs that have occurred between this time and the time it was registered as missing are considered as illegal accesses.

[2]Missing smart cards may of course also be misused by other people that are not health-care professionals, but in the case of constraint C_1 we only focus on health-care professionals.

It seems reasonable to monitor different types of violations of constraint C_1 in order to measure its degree of fulfillment. The violations of interest, for this particular constraint, are the different types of illegal accesses that may be performed by health-care professionals. These are as follows:

1) Not approved unauthorized accesses to EPRs where the owners of the EPRs are patients of the accessors
2) Not approved unauthorized accesses to EPRs where the owners of the EPRs are not patients of the accessors
3) Accesses to EPRs from missing or stolen smart cards

It should be noticed that each EPR is owned by a patient, which is natural since the information stored in the EPR is about the patient in question.

For each of the types 1, 2, and 3 of illegal accesses we identify the key indicators K_1, K_2, and K_3, respectively, where each key indicator measures the ratio of one type of illegal accesses to all accesses to information in EPRs. In (1) these key indicators have been used to express C_1. Ideally, Client H would have liked all three ratios to be zero in order for the constraint to be fulfilled. However, in real life things are not perfect; EPRs may for example be accessed by accident or a health-care professional may not have a perfect recollection of when he/she used the smart card the last time before losing it. After some hesitation, Client H came up with the intervals documented in (1).

$$0 \leq K_1 \leq 0.005 \ \wedge \ 0 \leq K_2 \leq 0.001 \ \wedge \ 0 \leq K_3 \leq 0.0001 \quad (1)$$

The formula in (1) expresses for what key indicator values constraint C_1 is fulfilled. By inserting key indicator values into this formula we get the degree of fulfillment of C_1. For instance, if K_1 equals 0.006 while the values of K_2 and K_3 are less than their upper thresholds, then C_1 is close to being fulfilled. On the other hand, if K_1 equals 0.05 instead, then C_1 is far from being fulfilled.

B. Specify key indicator requirements (Step 2.2 of ValidKI)

The key indicators identified in the previous step only provide a high-level specification of what they should measure. In Table I a more refined specification has been given for each key indicator. As we can see, each of the three identified key indicators is composed of basic key indicators. The expectation to each basic key indicator is that it is measured every week and that it measures the total number for all three health-care institutions. This is necessary since Client H is responsible for the handling of EPRs not only at its own premises but also within its two outsourcing partners when they act on behalf of Client H. In addition, we specify for each key indicator used in (1) the required level of trust in the correctness of its values. Together with Client H we

 ISBN: 978-1-61208-160-1

Table I
KEY INDICATOR REQUIREMENTS SPECIFICATIONS

The composite key indicator K_1 is the ratio of the basic key indicator K_{B1} = "the total number of not approved unauthorized accesses at Client H, Blood test analysis, and X-ray, in the period of one week, where the owners of the EPRs are patients of the accessors" to the basic key indicator K_{B2} = "the total number of accesses to EPRs at Client H, Blood test analysis, and X-ray in the period of one week". Required level of trust in the correctness of the values of K_1: 0.9
The composite key indicator K_2 is the ratio of the basic key indicator K_{B3} = "the total number of not approved unauthorized accesses at Client H, Blood test analysis, and X-ray, in the period of one week, where the owners of the EPRs are not patients of the accessors" to the basic key indicator K_{B2}. Required level of trust in the correctness of the values of K_2: 0.9
The composite key indicator K_3 is the ratio of the basic key indicator K_{B4} = "the total number of accesses at Client H, Blood test analysis, and X-ray, in the period of one week, from smart cards registered as missing" to the basic key indicator K_{B2}. Required level of trust in the correctness of the values of K_3: 0.9

assign a trust level of 0.9 to each composite key indicator. This means that for each composite key indicator we need to believe that the probability of its values being correct is at least 0.9, in order for the composite key indicator to be useful.

C. Evaluate external validity (Step 2.3 of ValidKI)

To evaluate external validity we try in collaboration with Client H to construct an argument for external validity based on the reformulated precise business objective and the requirements specifications of the key indicators. At this stage we assume for each key indicator that it is possible to come up with a realization that is internally valid. We agree that the identified key indicators can be used to monitor possible violations of constraint C_1, but we are a bit uncertain whether we have managed to capture all the types of illegal accesses that are relevant when measuring the degree of fulfillment of C_1. After some discussion, we conclude that there is fourth type of illegal accesses that we should also monitor. This type of illegal access may occur if a health-care professional forgets his/hers smart card in a terminal used to access information in EPRs. Other health-care professionals may then use this terminal to access information in EPRs. An indication of the number of forgotten smart cards is the number of terminals that timed out with a smart card inserted.

Table II
ADDITIONAL KEY INDICATOR REQUIREMENTS SPECIFICATION

The composite key indicator K_4 is the ratio of the basic key indicator K_{B5} = "the total number of smart cards at Client H, Blood test analysis, and X-ray, in the period of one week, left in terminals used to access information EPRs and where the terminal eventually timed out" to the basic key indicator K_{B6} = "the total number of smart cards used at Client H, Blood test analysis, and X-ray in the period of one week". Required level of trust in the correctness of the values of K_4: 0.9

We update the models created in Step 2.1 and 2.2 in order to include the new type of illegal accesses. The Boolean expression in (1) is updated as shown in (2). The key indicator K_4 is the ratio of the number of forgotten smart cards to the number of smart cards used by all health-care professionals. The interval that Client H came up with for K_4 is documented in (2). In Table II, Client H's expectations to K_4 has been documented.

$$0 \leq K_1 \leq 0.005 \ \wedge \ 0 \leq K_2 \leq 0.001 \ \wedge \\ 0 \leq K_3 \leq 0.0001 \ \wedge 0 \leq K_4 \leq 0.002 \quad (2)$$

VI. INTERNAL VALIDATION

Due to lack of space, we only show how the internal validity of the key indicator K_1 is established.

A. Specify key indicator designs (Step 3.1 of ValidKI)

Together with Client H we specify the designs of the key indicators $K_{B1}(X)$, $K_{B2}(X)$, and K_1 in the form of algorithms, as shown in Table III. $K_{B1}(X)$ and $K_{B2}(X)$ are computed at each of the three health-care institutions. The total sum of the three variants of $K_{B1}(X)$ is denoted by K_{B1}, while the total sum of the three variants of $K_{B2}(X)$ is denoted by K_{B2}. The algorithms for $K_{B1}(X)$, $K_{B2}(X)$ are used by all three health-care institutions, while the algorithm for K_1 is only used by Client H. This algorithm takes the three variants of both $K_{B1}(X)$ and $K_{B2}(X)$ as input.

B. Specify key indicator deployments (Step 3.2 of ValidKI)

Together with Client H we create deployment specifications for each of the three health-care institutions. Each specification describes how data on which the calculation of K_1 is based is extracted and transmitted. The deployment specification for X-ray in Table IV specifies how different data is extracted, how often, and by whom; how the data is transmitted internally at X-ray; and how the data is transmitted to Client H, to who, by whom, and how often. The

 ISBN: 978-1-61208-160-1

Table III
KEY INDICATOR DESIGN SPECIFICATIONS FOR $K_{B1}(X)$, $K_{B2}(X)$, AND K_1 IN THE FORM OF ALGORITHMS

Algorithm for $K_{B1}(X)$
Input: $L_{B1}(X)$ = "list of all the unauthorized accesses to EPRs in the period of one week at X, where the owners of the EPRs are patients of the accessors", where $X \in \{\text{Client H}, \text{Blood test analysis}, \text{X-ray}\}$
Step 1: At X a manual inspection of the elements in $L_{B1}(X)$ is done. The list elements are partioned into two lists; one list containing the approved unauthorized accesses and one list containing the not approved unauthorized accesses. The employee at X partioning $L_{B1}(X)$ decides whether an unauthorized access should be classified as approved or not approved.
Step 2: $K_{B1}(X)$ is calculated by counting the number of list elements in the list of not approved unauthorized accesses.
Output: $K_{B1}(X)$ = "number of not approved unauthorized accesses at X in the period of one week, where the owners of the EPRs are patients of the accessors"
Algorithm for $K_{B2}(X)$
Input: $L_{B2}(X)$ = "list of all the accesses to EPRs in the period of one week at X", where $X \in \{\text{Client H}, \text{Blood test analysis}, \text{X-ray}\}$
Step 1: An employee at X calculates $K_{B2}(X)$ by counting the number of list elements in $L_{B2}(X)$.
Output: $K_{B2}(X)$ = "number of accesses to EPRs at X in the period of one week"
Algorithm for K_1
Input: K_{B1}(Client H), K_{B1}(Blood test analysis), K_{B1}(X-ray), K_{B2}(Client H), K_{B2}(Blood test analysis), and K_{B2}(X-ray)
Step 1: Calculate K_{B1} as follows: $K_{B1} = K_{B1}(\text{Client H}) + K_{B1}(\text{Blood test analysis}) + K_{B1}(\text{X-ray})$
Step 2: Calculate K_{B2} as follows: $K_{B2} = K_{B2}(\text{Client H}) + K_{B2}(\text{Blood test analysis}) + K_{B2}(\text{X-ray})$
Step 3: Calculate K_1 as follows: $K_1 = \frac{K_{B1}}{K_{B2}}$
Output: K_1

deployment specification of Blood test analysis is similar, while the deployment specification of Client H differs from the other two with respect to how data is transmitted. Client H's deployment specification describes only internal data transmission.

C. Evaluate internal validity (Step 3.3 of ValidKI)

To evaluate the internal validity of K_1 we construct together with Client H an argument that its design and deployment specifications fulfills its requirements specification. After consulting both the requirements specification, in Table I, and the realization of K_1 we conclude that this is the case. The algorithm for K_1, in Table III, calculates the composite key indicator exactly as stated in its requirements specification. Also, all the data needed to calculate K_1 is specified to be extracted and transmitted every week at all three institutions. This is also in accordance with the requirements specification since it specifies that K_1 must be computed every week and with data from all three institutions. In addition, we discuss with Client H how much trust

Table IV
KEY INDICATOR DEPLOYMENT SPECIFICATION FOR X-RAY THAT DESCRIBES THE EXTRACTION AND TRANSMISSION OF DATA USED TO CALCULATE K_1

Extraction and transmission of L_{B1}**(X-ray)**
The EPR system administrator at X-ray creates the list L_{B1}(X-ray) every week, based on the access log of the EPR system at X-ray, which contains information on all accesses to information in EPRs at X-ray. The list is created by extracting all list elements in the access log that represents an unauthorized access where the owner of the EPR is a patient of the accessor and where the access occurred during the past seven days. The list is sent by encrypted email to the EPR monitoring officer at X-ray for further processing.
Extraction and transmission of L_{B2}**(X-ray)**
The EPR system administrator at X-ray creates the list L_{B1}(X-ray) every week, based on the access log of the EPR system at X-ray. The list is created by extracting all list elements in the access log that represents an access that occurred during the past seven days. The list is sent by encrypted email to the EPR monitoring officer at X-ray for further processing.
Transmission of K_{B1}**(X-ray) and** K_{B2}**(X-ray)**
The EPR monitoring officer at X-ray transmits K_{B1}(X-ray) and K_{B2}(X-ray) every week to the EPR monitoring officer at Client H by the use of an encrypted email.

we can have in the correctness of the different kinds of data used to compute K_1. During this discussion, Client H tells us that they have more trust in the correctness of the values of $K_{B1}(\text{X-ray})$ than in the values of $K_{B1}(\text{Blood test analysis})$, since they believe that the employees at X-ray are more competent than the employees at Blood test analysis in classifying unauthorized accesses. Also, Client H finds it unlikely that X-ray or Blood test analysis would provide them with manipulated data. Based on the discussion, we assign high trust levels to the different kinds of data used to compute K_1 and we combine the individual trust levels into a single trust level for K_1. We conclude that our trust in the correctness of the values of K_1 is at least 0.9.

VII. RELATED WORK

To the best of our knowledge, there exists no other method with sole focus on design of externally valid key indicators to monitor the fulfillment of business objectives. There is a tool framework called Mozart [8] that uses a model driven approach to create monitoring applications that uses key performance indicators. We do not focus on the implementation of key indicators, but we specify what is needed for implementing them. The work in [8] also differs from our work by not designing indicators from scratch, but by mining them from a data repository during the design cycle.

An important part of our method is the assessment of external validity of the key indicators we design. There exist other approaches that assess the validity of indicators in other contexts. For instance, in [9] measurement theory is used to validate the meaningfulness of IT security risk indicators. There are also examples of approaches that assess

 ISBN: 978-1-61208-160-1

the validity of specific sets of key indicators. For instance, in [10] the validity of indicators of firm technological capability is assessed, while the validity of indicators of patent value is assessed in [11].

There are several approaches that focus on measuring the achievement of goals. One example is COBIT [12], which is a framework for IT management and IT governance. The framework provides a IT governance model that helps in delivering value from IT and understanding and managing the risks associated with IT. In the governance model, business goals are aligned with IT goals, while metrics, in the form of leading and lagging indicators [13], and maturity models are used to measure the achievement of the IT goals. In our approach we do not focus on the value and risk that the use of IT has with respect to the business objectives. In our context, IT is relevant in the sense of providing the infrastructure necessary for monitoring the part of business that needs to comply with the business objectives.

Another way to measure the achievement of goals, is by the use of the Goal-Question-Metric [14], [15] (GQM) approach. Even though GQM originated as an approach for measuring achievement in software development, it can also be used in other contexts where the purpose is to measure achievement of goals. In GQM, business goals are used to drive the identification of measurement goals. These goals do not necessarily measure the fulfillment of the business goals, but they should always measure something that is of interest to the business. Each measurement goal is refined into questions, while metrics are defined for answering each question. No specific method, beyond reviews, is specified for validating whether the correct questions and metrics have been identified. The data provided by the metrics are interpreted and analyzed with respect to the measurement goal, to conclude whether it is achieved or not. One of the main differences between our method and GQM is that we characterize completely what it means to achieve a goal/objective. In GQM, however, this may be a question of interpretation.

In the literature, key indicators are mostly referred to in the context of measuring business performance. There exist numerous approaches for performance measurement. Some of these are presented in [16]. Regardless of the approach being used, the organization must translate their business objectives/goals into a set of key performance indicators in order to measure performance. An approach that is widely used [17] is balanced scorecard [3]. This approach translates the company's vision into four financial and non-financial perspectives. For each perspective a set of business objectives (strategic goals) and their corresponding key performance indicators are identified. However, the implementation of a balanced scorecard is not necessarily straight forward. In [18], Neely and Bourne identify several reasons for the failure of measurement initiatives such as balanced scorecards. One problem is that the identified measures do not measure fulfillment of the business objectives, while another problem is that measures are identified without putting much thought into how the data must be extracted in order to compute the measures. The first problem can be addressed in the external validation step of our method, while the second problem can be addressed in the internal validation step.

Much research has been done in the field of data quality. The problem of data quality is also recognized within the field of key indicators [19], [20]. In [21] a survey on how data quality initiatives are linked with organizational key performance indicators in Australian organizations is presented. This survey shows, amongst other things, that a number of organizations do not have data quality initiatives linked to their key indicators. Data quality should be taken into account when designing key indicators, since the use of key indicators based on poor quality data may lead to bad business decisions, which again may greatly harm the organization.

In [22], [23] the problem of key indicators computed from uncertain events is investigated. The motivation for this work is to understand the uncertainty of individual key indicators used in business intelligence. The authors use key indicators computed from data from multiple domains as examples. In the papers a model for expressing uncertainty is proposed, and a tool for visualizing the uncertain key indicators is presented.

VIII. Conclusion

The contribution of this paper is the new method *ValidKI* (Valid Key Indicators) for designing key indicators to monitor the fulfillment of business objectives. ValidKI facilitates the design of a set of key indicators that is externally valid with respect to a business objective, i.e. measures the degree to which the business or relevant part thereof complies with the business objective. To the best of our knowledge, there exists no other method with sole focus on design of externally valid key indicators to monitor the fulfillment of business objectives. The applicability of our method has been demonstrated by applying it on an example case addressing the use of electronic patient records in a hospital environment.

The demonstration of our method on the example case shows that the method facilitates the design and assessment of key indicators for the purpose of measuring the degree of fulfillment of business objectives. Even though ValidKI has been demonstrated on a realistic example case there is still a need to apply ValidKI in a real-world industrial setting in order to evaluate properly to what extent it has the characteristic mentioned above and to what extent it can be used to design key indicators for systems shared between many companies or organizations. By applying ValidKI in such a setting we will also determine to some extent whether it is time and resource efficient.

 ISBN: 978-1-61208-160-1

ValidKI is not restrictive when it comes to designing key indicators. The only restriction that ValidKI place on the design of key indicators is that it should be possible to realize them. This is a necessary restriction since a key indicator is of no value if it cannot be realized.

In the example case we have used trust levels to specify how much trust we need to have in the correctness of the different key indicators in order for them to be useful. In ValidKI it is up to the analysts to decide how to assess whether a key indicator has the necessary trust level or not. Thus, different approaches for reasoning about and aggregating trust can be applied for coming up with the trust level of a key indicator. As future work we will investigate the use of different approaches for reasoning about and aggregating trust in ValidKI.

ACKNOWLEDGMENTS

The research on which this paper reports has been carried out within the DIGIT project (180052/S10), funded by the Research Council of Norway, and the MASTER and NESSoS projects, both funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements FP7-216917 and FP7-256980, respectively.

REFERENCES

[1] Merriam-Webster Online Dictionary, "Definition of Indicator," http://www.merriam-webster.com/dictionary/indicator, 2008, Accessed: 2011-04-27 11:00AM CEST.

[2] Merriam-Webster Online Dictionary, "Definition of Indicates," http://www.merriam-webster.com/dictionary/indicates, 2008, Accessed: 2011-04-27 11:00AM CEST.

[3] R. S. Kaplan and D. P. Norton, "The Balanced Scorecard – Measures That Drive Performance," *Harvard Business Review*, vol. 70, no. 1, pp. 71–79, 1992.

[4] Object Management Group, "Unified Modeling Language Specification, version 2.0," 2004.

[5] Council of Europe, "Convention for the Protection of Human Rights and Fundamental Freedoms," 1954.

[6] European Court of Human Rights, "Press Release – Chamber Judgments 17.07.08," 17. July 2008.

[7] Helsedirektoratet, "Code of Conduct for Information Security – The Healthcare, Care, and Social Services Sector," http://www.helsedirektoratet.no/vp/multimedia/archive/00278/The_Code_of_conduct_278829a.pdf, 2. June 2010, Accessed: 2011-04-27 11:00AM CEST.

[8] M. Abe, J. Jeng, and Y. Li, "A Tool Framework for KPI Application Development," in *Proceedings of the IEEE International Conference on e-Business Engineering (ICEBE'07)*. IEEE Computer Society, 2007, pp. 22–29.

[9] A. Morali and R. Wieringa, "Towards Validating Risk Indicators Based on Measurement Theory," in *Proceedings of First International Workshop on Risk and Trust in Extended Enterprises*. IEEE Computer Society, 2010, pp. 443–447.

[10] T. Schoenecker and L. Swanson, "Indicators of Firm Technological Capability: Validity and Performance Implications," *IEEE Transactions on Engineering Management*, vol. 49, no. 1, pp. 36–44, 2002.

[11] M. Reitzig, "Improving Patent Valuations for Management Purposes – Validating New Indicators by Analyzing Application Rationales," *Research Policy*, vol. 33, no. 6-7, pp. 939–957, 2004.

[12] IT Governance Institute, "COBIT 4.1," 2007.

[13] W. Jansen, *Directions in Security Metrics Research*. DIANE Publishing, 2010.

[14] V. R. Basili and D. M. Weiss, "A Methodology for Collecting Valid Software Engineering Data," *IEEE Transactions on Software Engineering*, vol. SE-10, no. 6, pp. 728–738, 1984.

[15] R. V. Solingen and E. Berghout, *The Goal/Question/Metric method: A Practical Guide for Quality Improvement of Software Development*. McGraw-Hill International, 1999.

[16] A. Neely, J. Mills, K. Platts, H. Richards, M. Gregory, M. Bourne, and M. Kennerley, "Performance Measurement System Design: Developing and Testing a Process-based Approach," *International Journal of Operation & Production Management*, vol. 20, no. 10, pp. 1119–1145, 2000.

[17] T. Lester, "Measure for Measure," http://www.ft.com/cms/s/2/31e6b750-16e9-11d9-a89a-00000e2511c8.html#axzz1ImHJOLmg, 5. October 2004, Accessed: 2011-04-27 11:00AM CEST.

[18] A. Neely and M. Bourne, "Why Measurement Initiatives Fail," *Measuring Business Excellence*, vol. 4, no. 4, pp. 3–6, 2000.

[19] S. M. Bird, D. Cox, V. T. Farewell, H. Goldstein, T. Holt, and P. C. Smith, "Performance Indicators: Good, Bad, and Ugly," *Journal Of The Royal Statistical Society. Series A (Statistics in Society)*, vol. 168, no. 1, pp. 1–27, 2005.

[20] D. M. Eddy, "Performance Measurement: Problems and Solutions," *Health Affairs*, vol. 17, no. 4, pp. 7–25, 1998.

[21] V. Masayna, A. Koronios, J. Gao, "A Framework for the Development of the Business Case for the Introduction of Data Quality Program Linked to Corporate KPIs & Governance," in *Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO'09)*. IEEE Computer Society, 2009, pp. 230–235.

[22] C. Rodríguez, F. Daniel, F. Casati, and C. Cappiello, "Computing Uncertain Key Indicators from Uncertain Data," in *Proceedings of 14th International Conference on Information Quality (ICIQ'09)*. HPI/MIT, 2009, pp. 106–120.

[23] C. Rodríguez, F. Daniel, F. Casati, and C. Cappiello, "Toward Uncertain Business Intelligence: the Case of Key Indicators," *Internet Computing*, vol. 14, no. 4, pp. 32–40, 2010.

 ISBN: 978-1-61208-160-1