September 8th to 11th, 2015 - São Paulo - Brazil

http://ontobras.ime.usp.br

Sponsors

Organizing institutions

Editors addresses:

Fred Freitas

Universidade Federal de Pernambuco
Centro de Informtica
Av. Luz Freire, s/n
Cidade Universitria
50732-970 - Recife, PE - Brasil

Fernanda Baiao

Universidade Federal do Estado do Rio de Janeiro
Centro de Cincias Exatas e da Terra
Departamento de Informtica Aplicada
Av Pasteur 458 sala 114
Urca
22290240 - Rio de Janeiro, RJ - Brasil

Preface

This volume collects all papers accepted for presentation at Ontobras 2015, the VII Brazilian Ontology Research Seminar (7o Seminário de Pesquisa em Ontologias no Brasil), held on September 7-10, 2015 in Sao Paulo, Brazil. In its 7th edition, we sincerely hope Ontobras to be once more a valuable and pleasant opportunity for researchers and practitioners from Information Sciences and Computer Science to exchange their interesting ideas and experience on the exciting and broad field of Ontology, and to foster collaboration opportunities among research groups.

We received 32 submissions, from which 12 were accepted as full papers and 9 as short papers. All submissions were carefully evaluated by 3-4 reviewers, who provided excellent feedback for each work. The technical program of Ontobras2015 is also proud to include 3 keynote talks from distinguished researches: Professor Grigoris Antoniou, from the University of Huddersfield, UK; Dr. Enrico Franconi, from the Free University of Bozen-Bolzano, Italy; and Dr. Gilberto Câmara, from the National Institute for Space Research (INPE), Brazil.

We are very grateful to the authors for submitting their works, to all the program committee members for doing an excellent job, and to all the people involved in the local organization, represented by Prof. José M Parente de Oliveira, from ITA, Prof. Renata Wassermann, from USP; and Raphael Cóbe, from UNESP, for providing a concrete environment for this Seminar to occur.

The Ontobras 2015 Program
Committee Chairs
Fred Freitas
Fernanda Baiao

Table of Contents

Short Papers

Program Committee

Mara Abel	Universidade Federal do Rio Grande do Sul
Joao Paulo Almeida	Federal University of Espirito Santo
Mauricio Almeida	UFMG
Anarosa Alves Franco Brandão	University of Sao Paulo
Fernanda Baiao	UNIRIO
Monalessa Barcellos	UFES
Marcello Bax	Federal University of Minas Gerais
Regina Braga	UFJF
Maria Luiza Campos	PPGI - IM/NCE - Federal University of Rio de Janeiro
Daniela Claro	FORMAS/LASID/DCC/UFBA
Flavio S Correa Da Silva	Universidade de Sao Paulo
Evandro Costa	Federal University of Alagoas
Cedric de Carvalho	Universidade Federal de Goiás - Instituto de Informática
Alicia Diaz	Lifia, Fac. Informatica, UNLP
Frederico Durao	Federal University of Bahia
Jérôme Euzenat	INRIA & Univ. Grenoble
Ricardo A. Falbo	Federal University of Esprito Santo
Bernadette Farias Lóscio	Federal University of Pernambuco
Roberta Ferrario	Institute for Cognitive Sciences and Technologies - CNR
Renato Fileto	UFSC
Sandro Rama Fiorini	Universidade Federal do Rio Grande do Sul
Fred Freitas	Universidade Federal de Pernambuco (UFPE)
Renata Galante	UFRGS
Fernando Gauthier	Federal University of Santa Catarina/Engineering and Knowledge Management Graduate Program
Gustavo Giménez-Lugo	Federal University of Technology-Paraná (UTFPR)
Giancarlo Guizzardi	Ontology and Conceptual Modeling Research Group (NEMO)/Federal University of Espirito Santo (UFES)
Claudio Gutierrez	Universidad de Chile
Gabriela Henning	INTEC (CONICET-UNL)
Seiji Isotani	University of Sao Paulo
Fernanda Lima	UnB (Universidade de Brasilia)
Rinaldo Lima	UFPE
Lucelene Lopes	PUCRS
Andreia Malucelli	PUCPR
Carlos Marcondes	University Federal Fluminense
Ronaldo Mello	UFSC
Ana Maria Moura	LNCC

Alcione Oliveira	Universidade Federal de Viçosa
Emerson Paraiso	PUCPR - Pontificia Universidade Catolica do Parana
Jose M Parente De Oliveira	Aeronautics Institute of Technology
Osvaldo Pessoa	Universidade de São Paulo
Carlos Eduardo Pires	UFCG
Alexandre Rademaker	IBM Research Brazil and EMAp/FGV
Kate Revoredo	UNIRIO
Márcio Ribeiro	EACH - USP
Ryan Ribeiro De Azevedo	Federal University of Pernambuco
Rodrigo Rocha	UAG - UFRPE
Renato Rocha Souza	Fundação Getulio Vargas
Ana Carolina Salgado	Center for Informatics / UFPE
Lais Salvador	Computer Science Departament - UFBA
Stefan Schulz	Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz
Sean Siqueira	Federal University of the State of Rio de Janeiro (UNIRIO)
Damires Souza	IFPB
Cesar A. Tacla	CPGEI - UTFPR
Jose Todesco	Federal University of Santa Catarina/Knowledge Engineering and Management Graduate Program
Cassia Trojahn	UTM & IRIT
Marcela Vegetti	INGAR (CONICET / UTN)
Renata Vieira	PUCRS
Renata Wassermann	University of São Paulo

Additional Reviewers

Carbonera, Joel
Freire, Crishane
Oleynik, Michel

Using Ontology and Data Provenance to Improve Software Processes

Humberto L. O. Dalpra[1], Gabriella C. B. Costa[2], Tássio F. M. Sirqueira[1], Regina Braga[1], Cláudia M. L. Werner[2], Fernanda Campos[1], José Maria N. David[1]

[1]UFJF – Federal University of Juiz de Fora – Department of Computer Science, Juiz de Fora – MG – Brazil.

[2]UFRJ – Federal University of Rio de Janeiro – COPPE – Systems Engineering and Computer Science Department, Rio de Janeiro – RJ –Brazil.

humbertodalpra@gmail.com, tassio@tassio.eti.br,
{gabriellacbc,werner}@cos.ufrj.br,
{regina.braga, fernanda.campos, jose.david}@ufjf.edu.br

Abstract. *Provenance refers to the origin of a particular object. In computational terms, provenance is a historical record of the derivation of data that can help to understand the current record. In this context, this work presents a proposal for software processes improvement using a provenance data model and an ontology. This improvement can be obtained by process data execution analysis with an approach called PROV-Process, which uses a layer for storing process provenance and an ontology based on PROV-O.*

1. Introduction

Process can be defined as a systematic approach to create a product or to perform some task [Osterweil, 1987]. Currently, many organizations are investing in the definition and improvement of their processes aiming to improve product's quality. However, the increase of process data generated makes the analysis of them more complex. It requires the use of techniques to allow proper analysis of these data, extracting records that, in fact, will contribute to process improvement. One way of analyzing this data is using provenance techniques and models.

Buneman *et al*. (2001) define data provenance as the description of the origins of a piece of data and how it is stored in a database. Thus, to capture the origin of process data, it is necessary to capture the process flow specification (prospective provenance) and process execution data (retrospective provenance), in order to have the information regarding the success, failure, delays and errors, during process execution.

Lim *et al*. (2010) state that the provenance can be captured prospectively and retrospectively. Prospective provenance captures the abstract workflow specification (or process) enabling future data derivation. Retrospective provenance captures process execution, *i.e*., data derivation records.

To obtain the benefits of provenance, data have to be modeled, gathered, and stored for further queries [Marinho *et al*., 2012]. After the capture and storage of process provenance data, it can be used for analysis that enables process improvement (*e.g*., shorter execution time and greater efficiency of the results). One possible way to analyze processes provenance data is through the use of ontology and the inference mechanisms provided by it, enabling the discovery of strategic information for software project managers. This paper proposes a layer for the storage of software process

provenance data and the analysis of these data using an ontology. A W3C provenance model called PROV [Groth and Moreau, 2013] was used both for storage and analysis of these data.

The remainder of this paper is structured as follows: Section 2 presents related works that deal with provenance and processes. Section 3 is dedicated to describe the approach to improve software processes using an ontology called PROV-Process. The next section presents an overview of the PROV-Process ontology, which was based on PROV-O, describing the extensions made on it. Section 5 discusses the analysis of an industry software process using the PROV-Process approach and the possibilities to improve future executions of this process through the information obtained by PROV-Process ontology. Finally, conclusions are presented in Section 6.

2. Related Work

Missier *et al.* (2013) present D-PROV, an extension of PROV specification, with the aim of representing process structure, *i.e.*, to enable the storage and query using prospective provenance. An example of using D-PROV in the context of scientific workflows defined by Data ONE scientists was shown in the article. This work was used as basis to capture prospective provenance in PROV-Process approach.

Miles *et al*. (2011) propose a technique, called PRiME, to adapt application projects to interact with a provenance layer. The authors specify the steps involved in applying PRiME and analyze its effectiveness through two case studies.

Wendel *et al.* (2010) present a solution to failures in software development processes based on PRiME, the Open Provenance Model and a SOA architecture. They use Neo4j to store the data, Gremlin to query and REST web services as the connection to the tools.

Junaid *et al.* (2010) propose an approach where a provenance system intercepts the actions of users, processes and stores these actions to provide suggestions on possible future actions for the workflow project. These suggested actions are based on the actions of the current user and are calculated based on the provenance information stored.

Similar to the related work mentioned above, PROV-Process approach aims to improve future software process executions, through provenance data. However, other approaches do not use ontologies as a technique for query provenance data or use any inference mechanism, as PROV-Process approach does. Through ontology inferences, we derive strategic information to suggest software process improvement, as shown in the next sections.

3. PROV-Process Overview

PROV-Process is an approach for storage and analysis of software process provenance data in order to improve future process execution. The main objective of the approach is to identify improvements for future software process instances by using a provenance layer (comprising a database, an ontology and mechanisms to manipulate these components).

As shown in Figure 1, after the process modeling, a process instance can be created. Both the process model and the model of the generated instance are stored in

PROV-Process Database, through a prospective mechanism. After that, the process instance can be executed and the retrospective data provenance is stored through the PROV-Process approach. This storage is done using a relational database, which has been modeled using PROV-DM specification [Moreau and Missier, 2013].

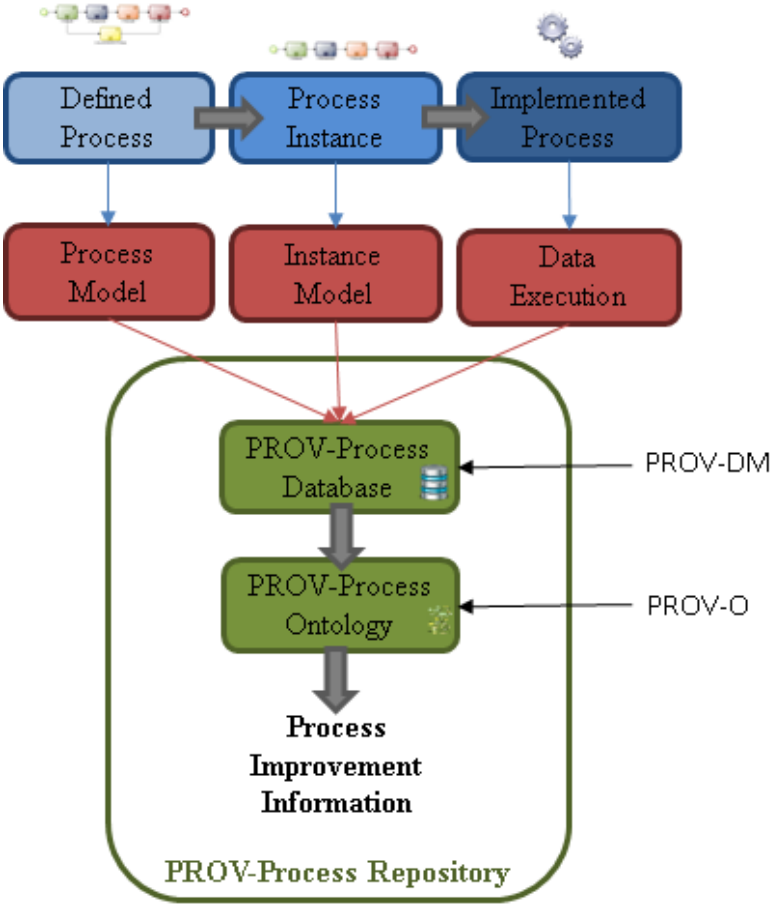

Figure 1: PROV-Process Approach

PROV-DM types and relations are organized according to six components. PROV-Process Database implements all these components using a relational database. Figure 2 shows, for example, tables of the first component, which comprise entities, activities and their interrelations: Used (Usage), WasGeneratedBy (Generation), WasStartedBy (Start), WasEndedBy (End), WasInvalidatedBy (Invalidation), and WasInformedBy (Communication).

All the data stored in the PROV-Process relational database are exported to the PROV-Process ontology. This ontology is described in details in next section.

4. PROV-Process Ontology

Ontology research has become more widespread in Computer Science community. Although the term has been limited to the philosophy sphere in the past, it has earned specific roles in Artificial Intelligence, Computational Linguistics and Databases [Guarino, 1998].

PROV-Process Ontology was developed from the PROV-O ontology [Belhajjame *et al*., 2013], which was defined based on PROV-DM data model. PROV-O defines the vertices of PROV (Agent, Entity and Activity) as classes and uses object properties for the interrelations representation. The core classes and properties from PROV-O are shown in Figure 3.

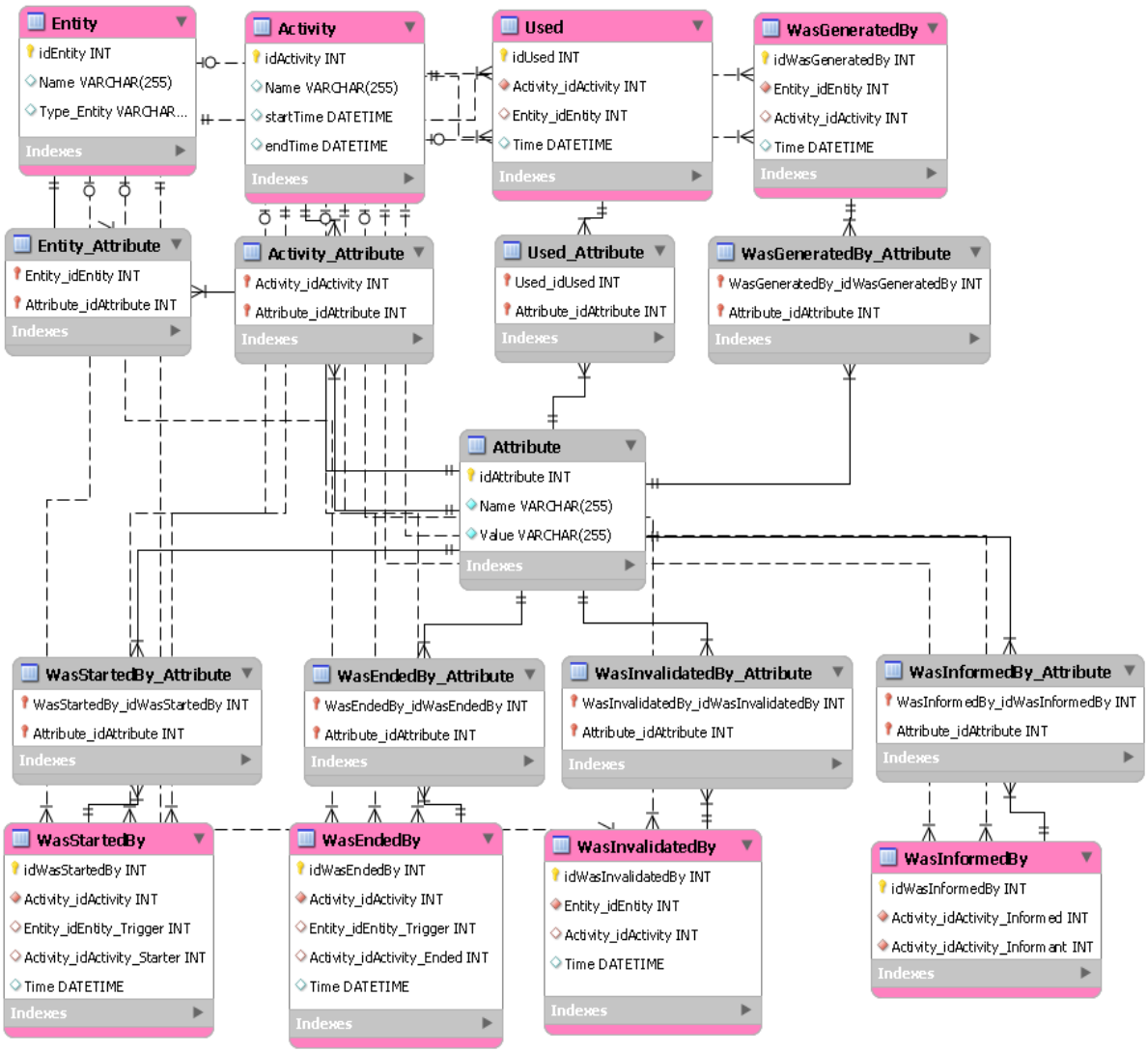

Figure 2: Part of PROV-Process Database

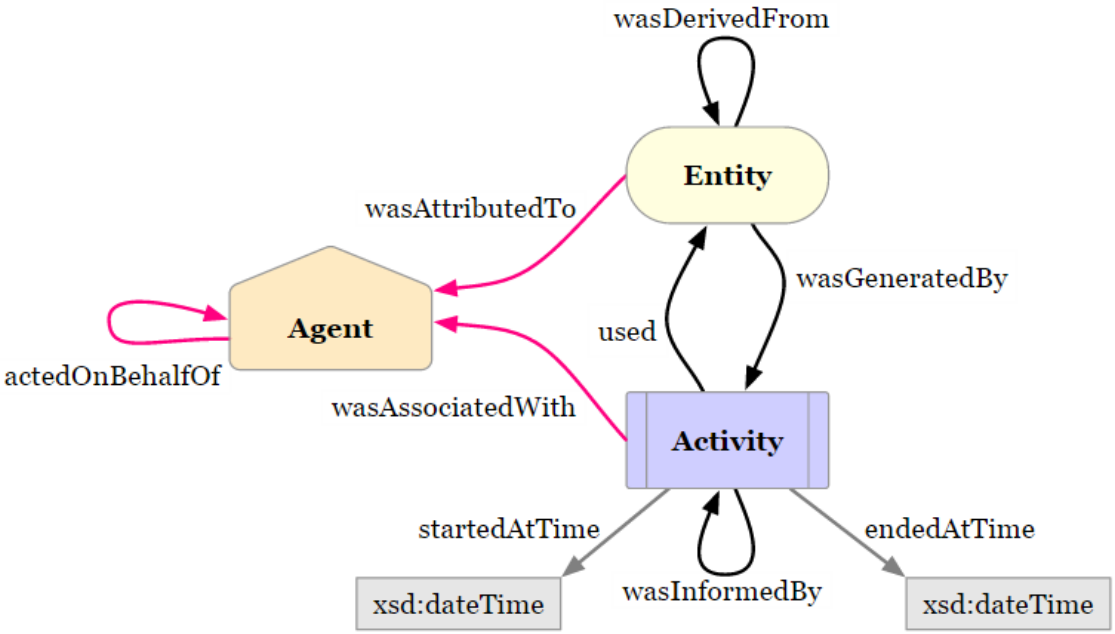

Figure 3: PROV-O: Core Classes and Properties [Belhajjame *et al.*, 2013]

Classes and properties in PROV-O can be used directly to represent provenance information or one can specialize them for modeling specific applications. Thus, PROV-O can also be specialized to create new classes and properties to model provenance information for different domains and applications. Based on this, we create some new properties on PROV-O (generating PROV-Process Ontology), in order to adapt it to the software process domain and to allow the inference of new information to improve software processes. Examples of these properties are presented in the following.

A group of rules (using Property Chains) was added in PROV-O in the 'wasAssociatedWith' data property:

1. **used** o **wasAttributedTo**
2. **wasStartedBy** o **wasAttributedTo**
3. **wasEndedBy** o **wasAttributedTo**

These rules state that, as show in Figure 4, if an activity used, was stated by or was ended by an entity and that entity was assigned to an agent, we can infer that an activity is associated with an agent.

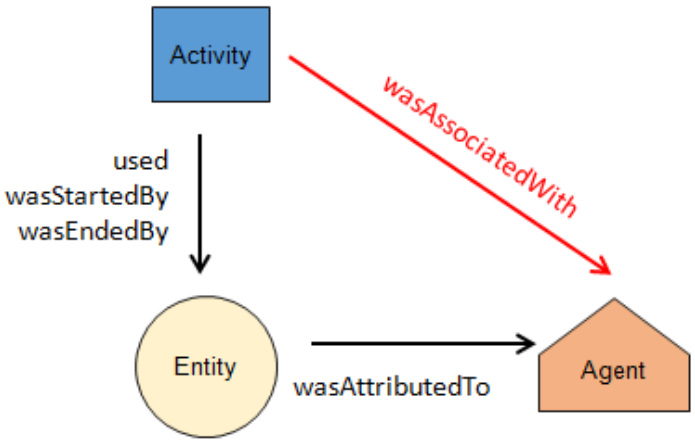

Figure 4: wasAssociatedWith properties chains

In the PROV-O, a data property called *processInstanceId* that corresponds to the generated/executed instance identifier from the main process was also inserted.

Finally, it should be noted that all records, called *Attributes* in PROV-Process database, must be exported to the PROV-Process Ontology as new data properties with their respective value.

5. Evaluation

In order to evaluate the applicability of the ontology of PROV-Process to software process, the approach was applied to a process from a Brazilian software development company [Ceosoftware, 2015]. A flow model shown in Figure 5 was created based on the specifications of this process.

To do this evaluation, real data execution of the process expressed in Figure 5 was analyzed. Thus, retrospective provenance of this process instances was stored using the PROV-Process relational database. It should be noted, however, that the execution data of the whole process were not provided by the company, but just a part of it.

In this work, 10 process execution instances, which have been fully completed, were analyzed. Regarding the obtained data, the following were used:

RDM[1] (change request) number;

- Information if an RDM was created from a previous RDM;
- Date and time of RDM opening;
- Type of RDM;
- Responsible for opening the RDM (Origin);
- Changed modules and components during the deployment task;
- Team responsible for implementation of the solution;

[1] RDM is an acronym for 'change request', in Portuguese, used by the company which provided the data for this research. It means a registration opened by support, client or commercial department, to make changes / adjustments in software system.

- Situation of RDM;
- Date and time of RDM completion.

These process execution data were obtained through a spreadsheet sent by the company responsible for the project[2].

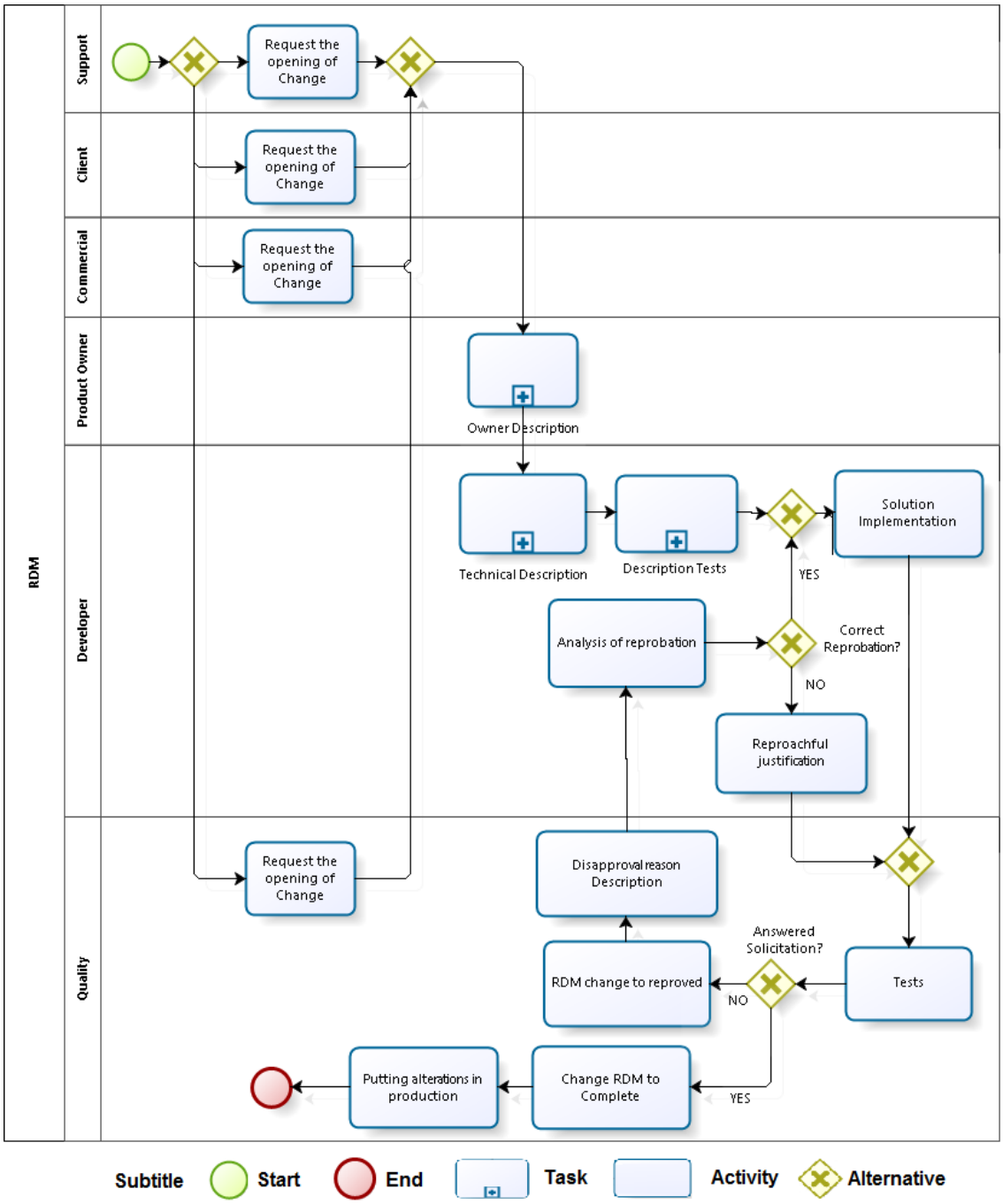

Figure 5: Process to manage requests and changes in software

[2] All the execution data used for the implementation of this assessment can be found at this link http://gabriellacastro.com.br/dsc/ex1/ex1.xlsx . Each row of this table represents a distinct execution of the process.

Table 1: Data execution example – Part 1

RDM Number	Outspread	Opening Date	Opening Time	Type	Origin
30006	0	10/03/2013	14:54:00	Module liberation	Client
30006	1	06/11/2014	17:18:00	Module liberation	Client

Table 2: Data execution example – Part 1

RDM Number	RDM Module	Module	Component	Team	Closed Date	Closed Time
30006	Financial	DLL - ERP PDA	clsValidacao	VB6	10/03/2013	22:06:00
30006	Financial	DLL - ERP PDA	clsValidacao	VB6	06/12/2014	10:41:00

The obtained data (examples about these data can be seen in Tables 1 and 2, where the dates are using the format MM/DD/YYYY) were imported to the PROV-Process database according to the following criteria:

- For all executions whose data were analyzed, three records were set in *Activity* table, with their names:
  - Opening the Request for Change;
  - Solution Implementation;
  - Change RDM to Complete.
- The RDM number was inserted as an attribute of each of the above activities, by using *Attribute* and *Activity_Attribute* tables.
- If a particular instance of execution corresponds to the unfolding of a previous RDM, a record in *WasInfomedBy* table was created.
- Date and time of RDM open were included using the *startTime* attribute of the activity Opening the Request for Change.
- RDM type was inserted as an attribute of the Opening the Request for Change activity using *Attribute* and *Activity_Attribute* tables.
- The role responsible for the Opening the Request for Change activity was inserted in *Agent* table, using the *name* field and the Person type.
- Relationship between Opening the Request for Change activity and the responsible for the same activity were inserted as records of the *WasAssociatedWith* table.
- Values as module, RDM module and component were included as records using *Entity* table and were associated with the Solution Implementation activity by creating records in *Used* table.
- Values as module, RDM module and component, included as records using *Entity* table, have been associated with agents who manipulated it by creating records in *WasAttributedTo* table.
- Role responsible for Solution Implementation activity was inserted in *Agent* table, using the *name* field and the Person type.

- Relationship between Solution Implementation activity and the responsible for the same activity were inserted as records in *WasAssociatedWith* table.
- Date and time of RDM completion were inserted using the *endTime* attribute of the Change RDM to Complete activity.
- As in the flow model (Figure 5) the role responsible for the Change RDM to Complete activity is the Quality Team. This role was inserted as record in the *Agent* table and was associated with this task by inserting a record in the *wasAssociatedWith* table.
- In order to identify which instance of the process execution a particular activity is associated with, a related attribute called *processInstanceId* was added to all activities by using the *Attribute* and *Activity_Attribute* tables.

After inserting the process execution data in the PROV-Process relational database, all the data were entered as individuals and their relationship in PROV-Process Ontology[3]. From this point, through the ontology inference engine, the derivation of strategic information was possible. As examples of information inferred from retrospective provenance data of this process, we can highlight four types:

1) Activities that influenced the generation of other activities, that is, as can be seen in red mark in Figure 6, Opening the Request for Change (id = 1) influenced Opening the Request for Change (id = 4). The same information was also inferred for the tasks of the same type with the ids 7, 13 and 19.

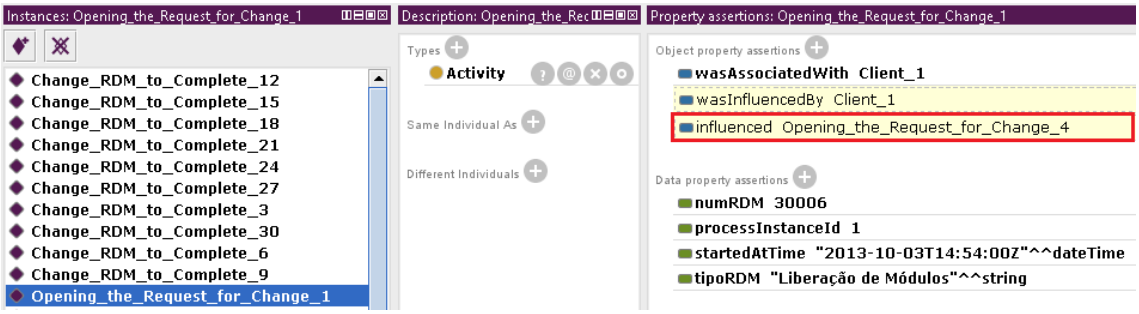

Figure 6: Activities that influenced the generation of other activities

2) Agents that could be associated with the Solution Implementation activity, considering that they already handled the artifacts involved in this activity in any other execution of the process. Figure 7 shows, for example, that Solution Implementation activity (id = 11) was influenced by DotNet agent (id = 5), given that this agent handled common artifacts to this activity in other instances of this process. The same type of information (agents that could be associated with the Solution deployment task) also occurs for Solution Implementation activity with ids equal to 8, 20, 23, 26 and 29.

[3] The generated ontology with all the individuals can be found at this link http://gabriellacastro.com.br/dsc/ex1/ex1-english.owl .

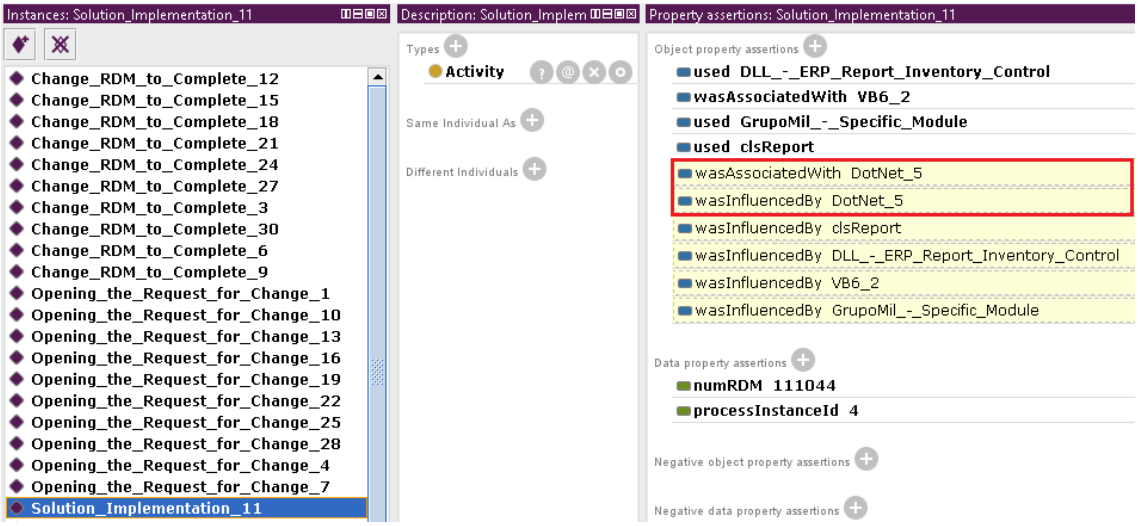

Figura 7: Agents that influenced an activity

3) A list of all activities in which an agent was involved, as well as the artifacts (entities) handled by her/him, as can be seen in Figure 8. Although this type of information can be obtained through queries on PROV-Process relational database, using the ontology and inference engine, this information can be obtained more easily (with a simple SPARQL query).

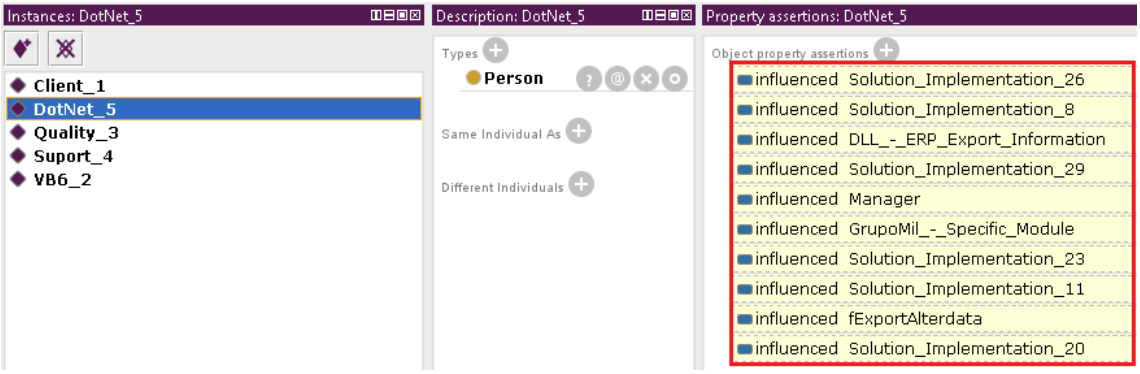

Figure 8: Activities and agents handled by DotNet agent

4) A list of all activities where an artifact (entity) was consumed, as can be seen in Figure 9. Although this type of information can be obtained through queries on PROV-Process relational database, using the ontology and inference engine, this information may be obtained more easily (with a simple SPARQL query).

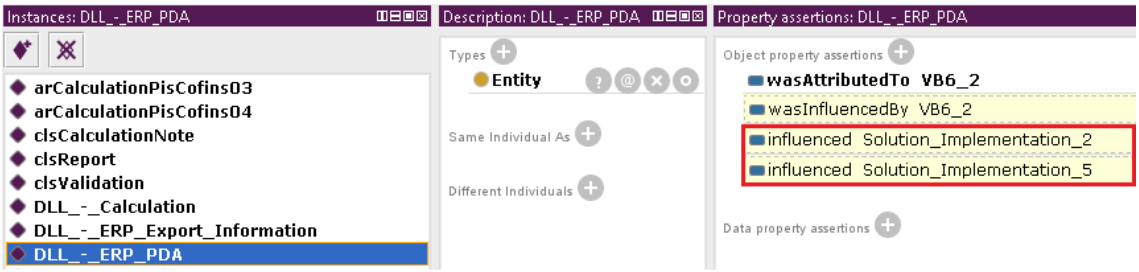

Figura 9: Activities where an artifact (entity) was consumed

Information inferred from the use of ontology proposed by the PROV-Process approach could help to improve process performance offering to the project manager information acquired at the time of the instantiation of a new process. This information might suggest, for example, the most appropriate agents and artifacts to be handled first, according to the type of problem reported / reason for opening the RDM.

6. Conclusion

This paper presented an approach, called PROV-Process, that obtains strategic information to the project manager enabling her/him to take decisions that can improve process performance. Therefore, this approach presents the advantages of using data provenance coupled with ontology. Through the use of ontology, it is possible to detect: (1) activities that influenced the generation of other activities; (2) agents that could be associated with the solution of the deployment task, considering that they already handled the artifacts involved in this task in any other execution of the process; (3) A list of activities in which an agent was involved, as well as the artifacts (entities) handled by her/him. No metric had be used for comparison of results.

Process data execution (retrospective provenance of the software process) are stored in a relational database, modeled based on PROV-DM specification. As a result, its data feed an ontology, created from the PROV-O model. With this and using an inference machine, one can infer new information about the process.

To evaluate the PROV-Process approach, it was applied to a real industry software process.

As threats to validity, we can cite:

- The partner company did not inform all process performance data, and only made available a spreadsheet with some of these data. This lack of detail directly impacts on a greater specificity in the results.
- Data obtained from the partner company did not include information about the actors who, in fact, performed the activity. They informed only the team that performed a certain activity.

Currently, we are working in the following improvements: (1) Implementing new rules indicating other actions that can help to improve processes; (2) Applying the approach to other real case studies.

References

Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J. (2013) "PROV-O: The PROV Ontology". Available in <http://www.w3.org/TR/prov-o/>. Accessed in July 2015.

Buneman, P., Khanna, S. and Tan, W. C. (2001) "Why and where: A characterization of data provenance". In: 8th International Conference on Database Theory, London. pp. 4-6.

Ceosoftware. (2015) "Soluções criativas e inovadoras". Available in <http://www.ceosoftware.com.br/>. Accessed in July 2015 (in Portuguese).

Groth, P., Moreau, L. (2013) "PROV - Overview". Available in <http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>. Accessed in July 2015.

Guarino, N. (1998) "Formal ontology in information systems" In: Proceedings of the first international conference (FOIS'98), Trento, Italy (Vol. 46). IOS press, pp. 3-15.

Junaid, M. M., Berger, M., Vitvar, T., Plankensteiner, K., Fahringer, T. (2009) "Workflow composition through design suggestions using design-time provenance information". In: E-Science Workshops, 2009 5th IEEE International Conference on. IEEE. pp. 110-117.

Lim, C., Lu, S., Chebotko, A., Fotouhi, F. (2010) "Prospective and Retrospective Provenance Collection in Scientific Workflow Environments". In Proceedings of the 2010 IEEE International Conference on Services Computing (SCC '10). IEEE Computer Society, Washington, DC, USA, pp. 449-456.

Marinho, A., Murta, L., Werner, C., Braganholo, V., Cruz, S. M. S. D., Ogasawara, E., Mattoso, M. (2012) "ProvManager: a provenance management system for scientific workflows". Concurrency and Computation: Practice and Experience, v. 24, n. 13, pp. 1513-1530.

Miles, S., Groth, P., Munroe, S., Moreau, L. (2011) "PrIMe: A methodology for developing provenance-aware applications". ACM Transactions on Software Engineering and Methodology (TOSEM), v.20 n.3, pp.1-42.

Missier, P., Dey, S. C., Belhajjame, K., Cuevas-Vicenttín, V., Ludäscher, B. (2013) "D-PROV: extending the PROV provenance model with workflow structure". In: Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP). USENIX Association, Berkeley, CA, USA, Article 9, pp. 1-7.

Moreau, L., Missier, P. (2013). "Prov-dm: The prov data model". Available in <http://www.w3.org/TR/2013/REC-prov-dm-20130430/>. Accessed in July 2015.

Osterweil, L. (1987) "Software processes are software too". In Proceedings of the 9th international conference on Software Engineering. IEEE Computer Society Press, pp. 2-13.

Wendel, H., Kunde, M., Schreiber, A. (2010). "Provenance of software development processes". In Provenance and Annotation of Data and Processes, v. 6378 of Lecture Notes in Computer Science, pp. 59-63. Springer Berlin Heidelberg.

An Empirical Study to validate the Use of Ontological Guidelines in the Creation of *i** Models

Ramilton Costa Gomes Júnior[1], Renata Silva Souza Guizzardi[1], Xavier Franch[2], Giancarlo Guizzardi[1], Roel Wieringa[3]

[1]Ontology and Conceptual Modeling Research Group (NEMO).
Federal University of Espírito Santo, Vitória/ES, Brasil.

[2]Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain

[3]University of Twente
Enschede, The Netherlands

ramiltoncosta@gmail.com,{rguizzardi,gguizzardi}@inf.ufes.br, franch@essi.upc.edu,roelw@cs.utwente.nl

Abstract. *i* is a well known goal modeling framework, developed by a large and geographically dispersed research community. Currently, i* users tend to ascribe different and conflicting meanings to its constructs, leading to a non-uniform use of the language, and consequently undermining its adoption. In previous works, we proposed ontological guidelines to support the creation of i* models, in an attempt to provide a solution to this problem. In this paper, we present an empirical study, to evaluate these ontological guidelines. Results show that for more experienced conceptual modelers, the ontological guidelines indeed support i* modeling. However, results are not as positive for non-experienced conceptual modelers.*

1. Introduction

*i** is a goal modeling framework used for Requirements Engineering (Yu, 1995). In the past twenty years, this framework has attracted the attention of different research groups, which have-proposed different variants of the initial framework, each one proposing different semantics to the language's constructs. The community that develops *i** is aware that this non-uniform use of *i** makes it difficult for novices to learn how to use the language, besides undermining its acceptance in industry.

We believe this problem can be solved with the use of a foundational ontology to interpret the semantics of the *i** concepts. A foundational ontology is a formal system of domain-independent categories that can be used to characterize the most general aspects of concepts and entities that belong to different domains in reality (Guizzardi, 2005). The idea is to apply the foundational ontology as a reference model to interpret the concepts of the language. Then, based on such interpretation, we are able to provide some guidelines to support modeling, here referred as ***ontological guidelines***. In previous works (Guizzardi, Franch, Guizzardi, 2012), (Guizzardi, Franch, Guizzardi, Wieringa, 2013), we proposed some ontological guidelines for *i** modeling, based on

the UFO foundational ontology (Guizzardi, 2005) (Guizzardi et al, 2013)(Guizzardi, Falbo, Guizzardi, 2008). The aim of this paper is to present the experimental design and the results of an empirical study conducted to evaluate the use of such ontological guidelines.

Nowadays, empirical studies are considered appropriate means to prove the effectiveness of a new approach. For (Vokac, 2002), the ideal science would have a set of empirical observations for each theory, either to support the theory or to prove it wrong. In other words, empirical observation is the core of the scientific process. Furthermore, it is through empirical observation that one can check theories, explore critical factors and give light to new phenomena, so that the theories can evolve (Travassos, 2002).

Having this in mind, we decided to conduct an experiment to confirm our intuitions that the use of ontological guidelines lead to *i** models of better quality. The experiment was conducted in two colleges, having fifty-five subjects in total. The subjects were students of a Systems Analysis and Development course and the PhD and Master program in Computer Science. The main goal of the experiment was to verify if the ontological guidelines cited above are useful or not in the development of *i** models. For that, the subjects participated in modeling activities with and without the use of the guidelines and, then, the results were compared. In the experiment applied with PhD and master students, the results show that the ontological guidelines are useful for the development of *i** models. Among the population of the second experiment application, composed of less experienced conceptual modelers, the experiment results were not so positive.

The remainder of this article is organized as follows: Section 2 presents information on the *i** framework and its variants; Section 3 describes the UFO fragment applied in this work; Section 4 presents some of the proposed ontological guidelines; Section 5 describes the empirical study; and, finally, Section 6 concludes the paper.

2. The *i** Framework and its Variants

The original *i** framework is described in (Yu, 1995). Since then, several variants have been proposed, for instance GRL and Tropos, see (Cares, 2012) for an overview. Some variants come from paradigm shifts, others propose some particular type of new construct, and still others issue slight modifications related to the core constructs of the *i** language.

One of the most controversial constructs in the language is the means-end link. In the original *i** (Yu, 1995), this link is used to connect a goal or a task to softgoals. In GRL, this link is applied to connect a task to a goal, a task to a task and a resource to a task. However, in the *i** wiki, one of the major sources of material about the language, this link is only used to connect a task to a goal. (Cares, 2012) also points out that different versions of Tropos propose different uses for the means-end link.

These different interpretations and uses make a new *i** learner confuse. She may ask herself: when can I use a means-end link after all? Why is it used this way? Why can't I use a means-end link between a resource and a goal, for example? We argue that

the best way to respond to these questions is to understand the ontological semantics behind the constructs of the language. By understanding their ontological nature, we may provide good reasons why a concept or a link may or may not be used in a particular way.

3. Background: The UFO Foundational Ontology

Here we briefly present the UFO concepts that are used in this paper provide an interpretation to *i**. To facilitate reading we use a different font to highlight the UFO concepts. For a fuller presentation on UFO, the reader should refer to (Guizzardi, 2005), (Guizzardi et al, 2013) and (Guizzardi, Falbo, Guizzardi, 2008).

In UFO, a stakeholder is represented by the `Agent` concept, defined as a concrete `Endurant` (i.e. an entity that endures in time while maintaining its identity) which can bear certain `Intentional States`. These intentional states include `Beliefs`, `Desires` and `Intentions`. `Intentions` are mental states of `Agents` which refer to (are about) certain `Situations` in reality. `Situation` are snapshots of reality. The `propositional-content` (i.e., proposition) of an `Intention` is termed a `Goal`.

In contrast to `Endurants`, `Events` are perduring entities, i.e., entities that occur in time accumulating their temporal parts. `Events` are triggered-by certain `Situations` in reality (termed their pre-situations) and they change the world by producing a different post-situation. `Action` are deliberate Events, i.e., `Events` deliberately performed by `Agents` in order to fulfill their `Intentions`. An `Action` achieves a `Goal` if the `Action` brings about a `Situation` in the world that satisfies that `Goal`.

In contrast with an `Agent`, an `Object` is a concrete `Endurant` that does not bear intentional states or perform actions. An `Object` participating in an `Action` is termed a `Resource`.

4. Ontological Guidelines for the Creation of *i** Models.

In this section, we describe some of the proposed ontological guidelines. For lack of space, we are not able to present them all and refer to (Guizzardi, Franch, Guizzardi, 2012) and (Guizzardi, Franch, Guizzardi, Wieringa, 2013) for a full description. In total, there are seven ontological guidelines and all of them have been considered in the experiment.

First, it is important to point out that we interpret *i** goals, tasks, resources and agents as their counterparts in UFO (with Action as task). Having that in mind, let us try to interpret the *i** decomposition relation. Since goals are propositions, due to its ontological nature, it is impossible for a goal to be decomposed into tasks or resources. Thus, goals can only be decomposed into subgoals. Consequently, when decomposing goals, an *i* and-decomposition* is interpreted as a conjunction of subgoals, while an *i* or-decomposition* is interpreted as a disjunction of subgoals. Similarly, softgoals, tasks and resources can only be decomposed into softgoal, tasks and resources, respectively. This originates the ontological guideline describe in the first line of Table 1.

In *i**, a means-end link is applied to connect a means to an end. For example, a task T (means) to a goal G (end), meaning that the execution of T leads to the achievement of G. Here, we adopt the conceptual modeling evaluation method proposed in (Guizzardi, 2005) that states that we should avoid construct redundancy, i.e., two language constructs should not be applied to model the same phenomenon in the world. Construct redundancy adds unnecessary complexity to the modeling language, besides making specifications more difficult to understand. Moreover, when facing redundancy, designers tend to ascribe slightly different meanings to the redundant constructs, which may not be fully understood by the model readers. In our case, if we allow, for instance, goals G2 and G3 to be connected via means-end to goal G1, we will not be able to differentiate between means-end and or-decomposition, i.e. these two links will be applied to represent the very same relation in the world. Thus, this will be a case of construct redundancy. To avoid that, we propose the ontological guideline described in the second line of Table 1.

In *i**, a make-contribution is applied between a task T and a goal G, meaning that if T is executed, then G is fully achieved. But if this is so, how can one differentiate between means-end and make-contribution? Using UFO, we differentiate this by looking at the intention behind the execution of T. To understand this, let us consider the *i** model depicted in Figure 1, which exemplifies the use of the means-end and the make contribution links.

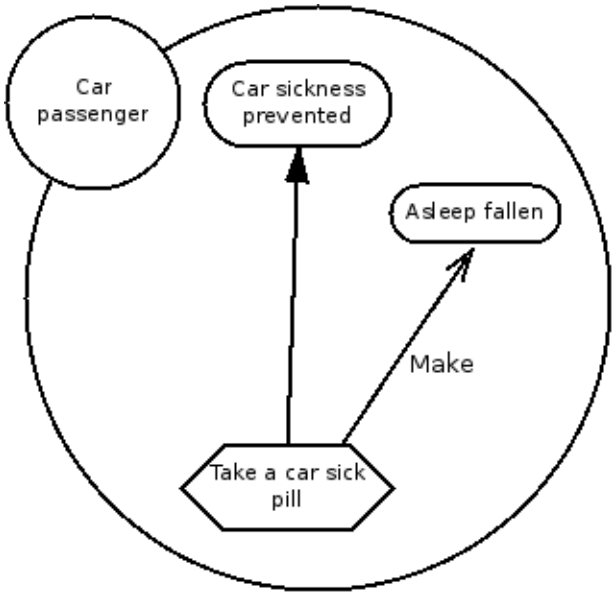

Figure 1. Means-end vs. make-contribution

In Figure 1, a Car Passenger[1] agent executes the Take a car sick pill task in order to prevent himself from being sick during the journey he is making (means-end link to Car sickness prevented goal). As a side effect of this medication, the Car Passenger also goes to sleep (make-contribution link to Asleep fallen goal).

As result of the mapping from *i** tasks into UFO actions, every task is associated with a motivating intention whose propositional content is a goal. In other words, we

1 From now on, we use a different font for the names of the instances of the *i** actors and intentional concepts, such as goals, tasks, and resources.

execute a particular task in order to accomplish a specific goal. In *i**, the association between the task and the goal in this case is made by a means-end link (e.g. Take a car sick pill task as means to Car sickness prevented goal). On the other hand, this same task can also generate some other goals to be accomplished, without however, being intended be the choice of this particular task. In this case, a make-contribution link is established (e.g. Take a car sick pill task contributing to asleep fallen goal). In other words, the means-end link or the make-contribution link should be applied according to the ontological guideline described in the third line of Table1.

Table 1. Some of the proposed *i ontological guidelines**

Ontological Guidelines
1. A decomposition link can only be applied between elements of *the same kind*. E.g. goal->goal, task->task.
2. A means-end link can only be applied between elements of *different kinds*. E.g. task->goal, resource->task.
3. Taking task T and goal G, if the intention behind the execution of task T is to accomplish G, T and G should be related via *means-end link*. On the other hand, if by executing T, G is unintentionally achieved (i.e., as a side-effect of the execution of T), then T and G should be related via *make-contribution*.

5. The Empirical Study

In this section, we describe the empirical study we conducted to evaluate the use of the ontological guidelines. The hypothesis of the study is *"the ontological guidelines enhance the capability of the subjects to create i* models."* The experiment was conducted in a controlled environment and is based on a quantitative strategy, in which the data is analyzed using statistical and descriptive methods. For the experimental design, we followed the framework presented in (Kochanski, 2009).

5.1 Experimental Design

The experiment has as object of study two *i** models (here referred to as Case 1 and Case 2), representing two different situations. Each participant had to complete the models, by filling in the blanks with the correct element or link to be used in each question. Figure 2 illustrates part of one model. For each blank, there are two and more possibilities, having as alternatives constructs of *i** whose use normally generates confusion or doubts. For example, in Question 2 (refer to Figure 2), the participants should indicate if “Provide gift wrapping solution” is a goal or a plan. In Question 5, the participants should indicate if “Provide gift wrapping solution” and the two tasks “Organize wrapping stand” and “Allow vendors to wrap gifts” should be linked via OR-means-end or via OR-decomposition. The idea is to verify if the participants can select them intuitively (pre-test) or if the use of ontological guidelines (post-test) effectively helps the selection of the correct construct.

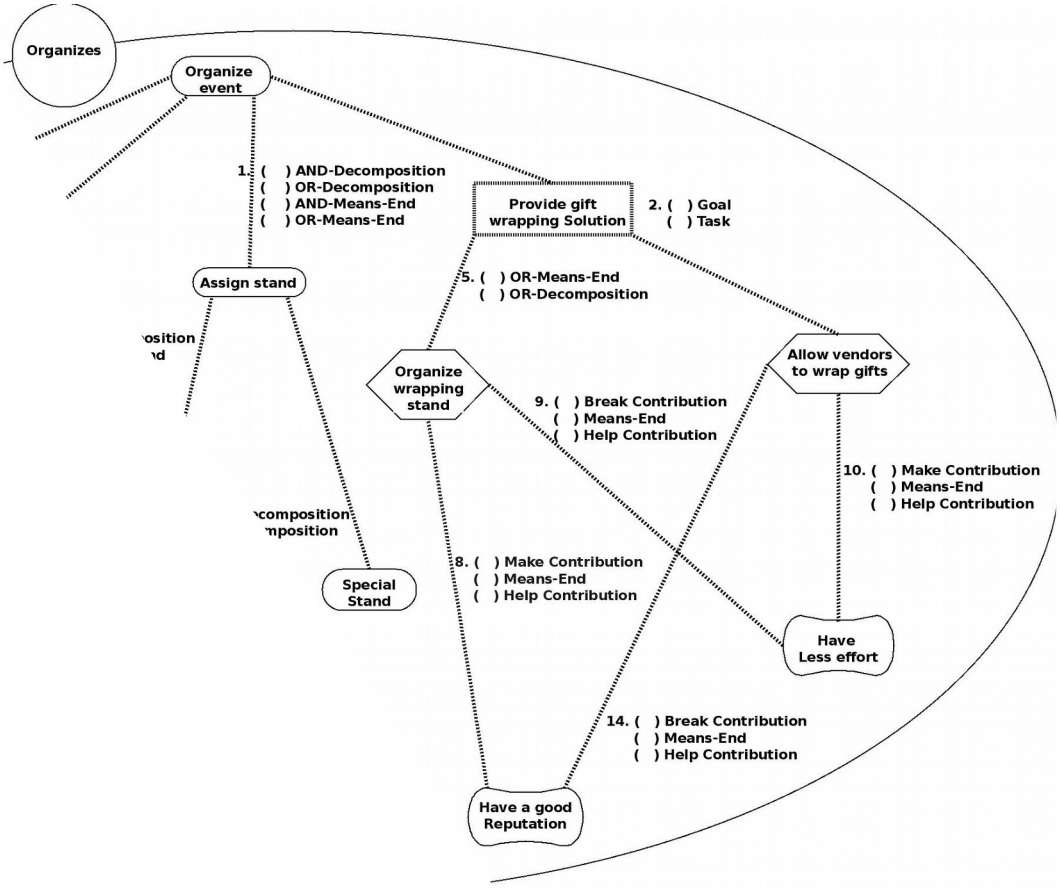

Figure 2. Part of the *i model**

The experiment was divided in two steps: pre-test and post-test. In the pre-test, all participants performed the first activity, i.e. filling in the blanks, using Case 1. Then, in a separate form, they justified their choices for each blank. During this activity, all participants had a printout of some slides containing basic information about *i** (the *i** wiki guidelines), as well as the description of Case 1. No information about the guidelines is given in this first step.

After the pre-test activity, the students were randomly divided into two groups: group A (control group) and group B (experimental group). After the division, the participants of group A moved to another room to perform the post-test activity. Both groups had to perform a second activity of filling in the blanks, now using Case 2. However, in this part, only group B received information about the ontological guidelines. Both groups had the description of Case 2 and group B also had a printout of some slides containing the ontological guidelines. In the post-test, the participants of both groups were also asked to fill in a separate form justifying their choices for each blank.

To capture the impression of the participants about the guidelines, the participants were also asked to respond some questions regarding their opinion about the *i** wiki guidelines and the ontological guidelines.

5.2 Collected Data

The data was collected through questionnaires. Before the experiment activities, we applied a questionnaire to capture the participants' profile. We applied the experiment twice, with two different populations. We will here refer to these applications as application 1 and application 2. In application 1, there were 24 participants: 16 of them were undergraduate students of Computer Science or Computer Engineering, 7 of them were master students in Computer Science, and 1 of them was a PhD student in Computer Science. The participants were assigned into two groups of 12 participants, which were balanced in terms of educational level and modeling experience. In both groups, there was one participant with 1-3 years of experience in goal modeling and *i**, while the others declared not having experience in this area. In application 2, there were 30 participants, all of them in the final year of an undergraduate course in Information Systems Analysis and Development. Each group had 15 participants. None of the participants indicated having experience in goal modeling or *i**.

Both in the pre-test and in the post-test, the same activities and questionnaires were used in applications 1 and 2. The graphs of Figures 3 and 4 show the results for the first and the second application, respectively. When the participant fills in the blank correctly, we say that he has a hit.

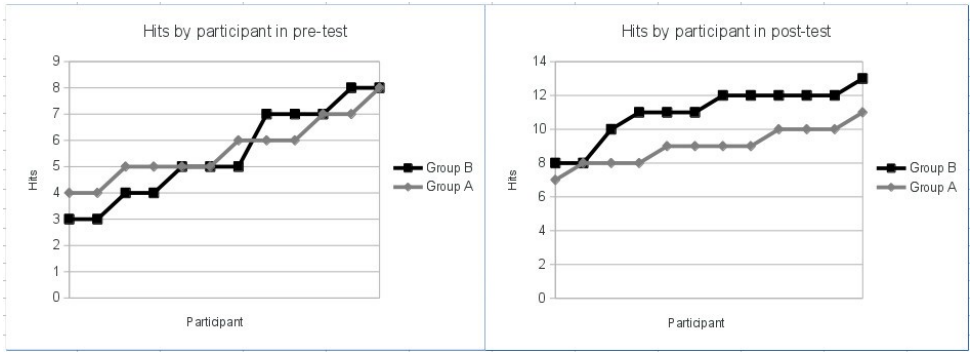

Figure 3. Hits by participant in pre-test (left) and post-test (right) in the first experiment application.

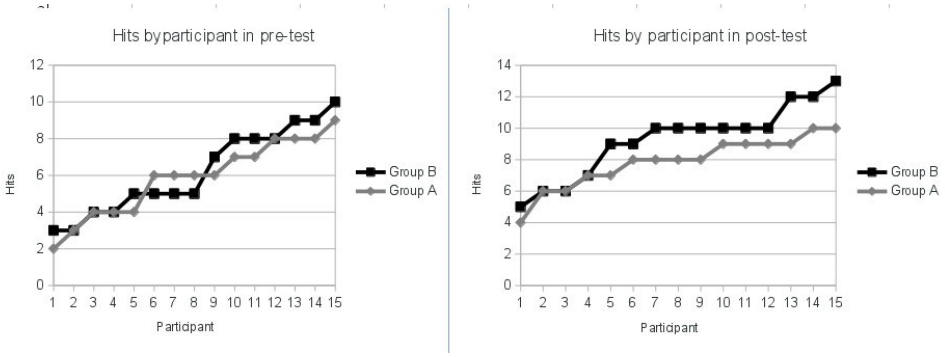

Figure 4. Hits by participant in pre-test (left) and post-test (right) the second experiment application.

Tables 2 and 3 present data regarding the number of hits per participant in the first and second application, respectively. The columns present data on average, median, highest and lowest value of number hits per participants.

Table 2 - Number of hits per participants in the first application

	Average		Median		Highest		Lowest	
	Group A	Group B	Group A	Group B	Group A	Group B	Group A	Group B
Pre-test	6,67	5,50	5,50	5,00	8,00	8,00	4,00	3,00
Post-test	9,00	11,00	9,00	11,50	11,00	13,00	7,00	8,00

Table 3 - Number of hits per participants in the second application

	Average		Median		Highest		Lowest	
	Group A	Group B	Group A	Group B	Group A	Group B	Group A	Group B
Pre-test	5,87	6,20	6,00	6,00	5,00	10,00	2,00	3,00
Post-test	7,89	9,27	9,27	8,00	10,00	13,00	4,00	5,00

5.3 Data Analysis

Analyzing Figure 3, we notice that in the pre-test of the first application, the participants of group A scored a larger number of hits than the participants of group B. However, in the post-test, group B performed better than group A. This shows that the group that used the ontological guidelines performed better when compared to the group that only had access to the *i** wiki guidelines. This result favors our hypothesis, supporting the idea that the ontological guidelines effectively help the creation of *i** models.

By looking at Figure 4, we see that in the pre-test of the second application, groups A and B showed a great balance in realizing the activities; both groups scored the same number of hits and errors. In the post-test, group B achieved a significantly higher number of hits in relation to group A, as seen in Figure 4. Again, this result favors our hypothesis, supporting the idea that the ontological guidelines effectively help the creation of *i** models.

Table 2 shows the data regarding the number of hits per participants in the pre-test and post-test, in the first application. The values for average, median, highest and lowest are very similar in the pre-test activity. But in the post-test activity, the values are significantly different, result that favors ours hypothesis.

Table 3 presents the data regarding number of hits per participants in the pre-test and post-test, in the second application. The values for average, median, highest and lowest have a small difference in the pre-test activity. But in the post-test activity, the values are significantly different, result that favors ours hypothesis.

The descriptive analysis we presented so far is able to provide us with some evidence supporting the hypothesis, We can quantify this support by a statistical test. Thus, we also applied the Wilcoxon-Mann-Whitney statistical test, with a significance level of 5%, to compare the hits for each participant between the experimental (group

B) and control (group A) groups, in both experiment applications. This statistical method is a non-parametric method recommended for small samples or groups with less than 20 participants (Robson 2002). In the first application, the calculated U value is 23 and the critical U value from the Mann-Whitney index is 37. Since the calculated U is lower than the critical U, then we may conclude that the values are significantly different between the groups, which supports our hypothesis. In the second application, the calculated U value is 65 and the critical U value from the Mann-Whitney index is 64. Since in this case, the calculated U is not lower than the critical U, then we cannot confirm our hypothesis.

Given the results of the Mann-Whitney test, we cannot conclude that the ontological guidelines are always helpful. We attribute this difference to the divergence in profiles in the two experiment applications. The participants of the first application have a higher graduation level than the participants of the second application, and thus are, in general, more experienced in conceptual modeling. Thus, we claim that the ontological guidelines are helpful for more mature conceptual modelers. New empirical studies should be conducted to confirm this hypothesis.

Regarding the qualitative evaluation of the ontological guidelines, we have the following results. In the first application, 7 out of 12 participants considered that the ontological guidelines are better than the *i** wiki guidelines. The other 5 participants considered that the ontological guidelines and the *i** wiki guidelines have the same quality. When asked about the usefulness of the ontological guidelines, 8 participants considered them very useful, 2 participants found them not very useful and 2 participants found them indifferent. In the second application, 13 out of 15 participants considered that the ontological guidelines are better than the *i** wiki guidelines, while 2 participants considered that the ontological guidelines and the *i** wiki guidelines have the same quality. Regarding the usefulness of the ontological guidelines, 10 participants found them very useful, 3 participants found them not so useful and 2 found them indifferent. We find these results positive, as most of the participants had a good perception regarding the ontological guidelines.

Let us now analyze which questions were more difficult, i.e. led to more errors in both experiment applications. This will allow us to find out which ontological guidelines are not clear and should be improved. In the first application, the questions that led to more errors were questions 8 and 10. In the second application, the questions that led to more errors were questions 7, 9 and 14. Questions 8, 9, 10 and 14 regard the use of the means-end, make-contribution and help-contribution links. We conclude that the participants in both experiment applications could not understand well the ontological difference between these three links. Thus, the ontological guidelines concerning this differentiation should be improved. Question 7 regards the differentiation among AND and OR decomposition. We conclude that in the second application, the participants also had doubts regarding the use of decomposition. Thus, the guidelines concerning these links should also be improved.

5.4 Threats to Validity

The following factors are considered the main threats to the validity of this empirical

study:

a) the heterogeneity of the participants of the first application, since they had different academics degrees. To mitigate this risk, we collected information about the academic degree of the participants in the profile questionnaire and took this into account in our experiment design;

b) the possibility that the participants had previous knowledge of the ontological guidelines. To remediate this risk, we asked in the experiment questionnaire if the participant had had previous contact with the guidelines. This information was taken into account in our analysis;

c) the chance that the participants had low interest in the experiment results, carelessly performing the experiment activities. To mitigate this risk, we tried to motivate the participants, showing the importance of the results of the experiment. Moreover, the experiment was designed to be as short as possible, so as to prevent tiredness and disinterest;

d) the possibility that the researcher conducting the experiment influenced the experiment results. To remediate this risk, the researcher conducting the experiment tried to be as objective and unbiased as possible during the experiment activities;

e) the possibility that the subjects had a positive opinion about the guidelines, because they knew we were the ones who formulated them. To remediate this risk, we did not tell them we were the authors of the guidelines.

6. Final Considerations

This article presented an empirical study with the objective to evaluate the use of ontological guidelines to create *i** models. For that, the experiment was conducted in two steps (pre and post-test), in which the participants performed modeling activities without (pre-test) and with (post-test) the use of ontological guidelines. To analyze the results, we performed the Mann-Whitney statistical test. The outcome supports our hypothesis that states that the guidelines are useful, and does not provide evidence against it. Moreover, most participants stated that they found the ontological guidelines useful to support them in the creation of *i** models.

Given the results of this experiment, we intend to develop an *i** modeling tool that uses the ontological guidelines as support for the model designer. For that, we aim at proposing a metamodel that is compatible with these guidelines, to serve as basis for the development of the tool.

For the future, we also intend to perform new experiments to collect more data regarding the use of the ontological guidelines to create *i** models. In order to confirm our hypothesis, we must repeat the designed experiment, taking populations of different profiles. We aim, for example, to conduct the experiment with professional modelers. Moreover, we intend to perform different experiments. For instance, we would like to conduct an experiment in which the participants are asked to create *i** models from scratch, with and without the use of the ontological guidelines. Then, based on some pre-established criteria collected from *i** experts, we will be able to analyze if the

models created with the use of ontological guidelines have higher quality than the ones created without them.

Acknowledgement. This work is partially supported by CAPES/CNPq (grant number 402991/2012-5), CNPq (grant numbers 461777/2014-2 and 485368/2013-7), and the Spanish project EOSSAC, ref. TIN2013-44641-P.

References

Ayala, C., Cares, C., Carvallo, J.P., Grau, G., Haya, M., Salazar, G., Franch, X., Mayol, E. and Quer, C. (2005), "A Comparative Analysis of *i**-Based Agent-Oriented Modeling Languages", In: 17th International Conference on Software Engineering and Knowledge Engineering, Taipei, Taiwan, pp. 43-50.

Cares, C. (2012), "From the *i** Diversity to a common interoperability framework", PhD Thesis, Software Engineering for Information System Research Group, UPC, Spain.

Guizzardi, G. (2005), "Ontological Foundations for Structural Conceptual Models". PhD Thesis, University of Twente, The Netherlands.

Guizzardi, G., Wagner, G., Falbo, R.A., Guizzardi, R.S.S., Almeida, J.P.A. (2013), Towards Ontological Foundations for the Conceptual Modeling of Events, 32nd International Conference on Conceptual Modeling (ER 2013), Hong Kong.

Guizzardi, G., Falbo, R. A., Guizzardi, R. S. S. (2008), Grounding Software Domain Ontologies in the Unified Foundational Ontology (UFO): The case of the ODE Software Process Ontology , 11th Iberoamerican Conference of Software Engineering (CIbSE 2008), Recife, 2008.

Guizzardi, R., Franch, X. and Guizzardi, G. (2012), "Applying a Foundational Ontology to Analyze Means-end Links in the *i** Framework", In: 6th IEEE International Conference on Research Challenges in Information Science, Valencia Spain, pp. 1-11.

Guizzardi, R., Franch, X., Guizzardi, G. and Wieringa, R. (2013), "Ontological Distinctions between Means-end and Contribution Links in the *i** Framework", Lecture Notes in Computer Science, v. 8217, Heidelberg: Springer, pp. 463-470.

Kochanski, D. (2009), "Um Framework para Apoiar a Construção de Experimentos na Avaliação Empírica de Jogos Educacionais", Master Dissertation in Applied Computing, UNIVALI, Brazil.

Lucena, R., Santos, B., ; Silva, J., Silva, L., Alencar, R. and Castro, B. (2008), "Towards a Unified Metamodel for *i**". In: 2nd IEEE International Conference on Research Challenges in Information Science, Marrakech, v. 1, pp. 237-246.

Santos, B. (2008), "Istar Tool – Uma proposta de ferramentas para Modelagem de *i**", Master Dissertation in Computer Science, UFPE, Brazil.

Travassos, G. (2002), "Relatório Técnico RT-ES-590/02 – Introdução à Engenharia de Software Experimental". Systems Engineering and Computer Science Program. COPPE/UFRJ, Brazil.

Vokac, M. (2002), “Empiricism in Software Engineering: A Lost Cause?” Essay for MNVIT401.

Yu, E. (1995), “Modelling Strategic Relationships for Business Process Reengineering”, Ph.D. thesis, Dept. of Computer Science, University of Toronto, Canada.

Yu, E. (1997), “Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering”, In: 3rd IEEE International Symposium on Requirements Engineering. Annapolis, USA, pp. 226-235.

Exploring Ontologies for Semantic Documentation in Project Management

Erick Casagrande Bastos, Monalessa Perini Barcellos, Ricardo de Almeida Falbo

Ontology and Conceptual Modeling Research Group (NEMO), Department of Computer Science, Federal University of Espírito Santo– Vitória – ES – Brazil

{erickcasagrande, monalessa, falbo}@inf.ufes.br

Abstract. *Although there are several tools devoted to support project management, documents are widely used as an instrument to record information regarding projects. However, retrieving information from documents is usually not trivial and depends on human effort. In this paper we discuss the use of semantic annotation of desktop documents in the project management context. The main results of a study that investigated initiatives involving semantic annotation to support project management aspects are presented, as well as an ongoing work in which we explore a software project management domain ontology to annotate desktop documents and extend a semantic document management platform.*

1. Introduction

Documents are an important instrument to record and share information in the project management domain, since they provide useful information for communication between people and for an effective understanding about the project [Bruggemann et al. 2000].

There are several tools to support project management, but they are not used by all organizations. Spreadsheets are widely used for organizations that have limited access to sophisticated tools to support some project management activities, such as schedule and budget planning and control [Villalobos et al. 2011]. Furthermore, project management supporting tools often do not eliminate the need of using desktop documents (e.g., text documents and spreadsheets).

One disadvantage of using documents is the difficulty of obtaining consolidated information from them. The access to their contents typically depends on human intervention, since they were originally designed to be read by humans, not to be manipulated by machines. As a consequence, retrieving and analyzing document content can be unproductive and sometimes inefficient. Besides, gathering relevant information from different documents can be so wearing that people may tend not to do that [Arantes and Falbo 2010].

In the Semantic Web community, researchers have defended that ontology-based metadata can be added into web contents so that these contents become available for machine interpretation. The act of adding ontology-based metadata into syntactic information resources making them semantic information resources is named *semantic annotation*. Ontologies are an ideal vehicle for describing the vocabulary for metadata statements, providing a rich formal semantic structure for their interpretation. Therefore ontology is often used as basis for annotation [Sicilia 2006].

Semantic Web principles can be applied to documents rendered by desktop tools (e.g., text and spreadsheet editors), giving rise to *Semantic Documentation*, which aims at making document content interpretable by computers. In this context, several tools have been developed to support semantic annotation, such as the Infrastructure for Managing Semantic Documents (IMSD) [Arantes and Falbo 2010], PDFTab [Eriksson 2007] and KIM [Kiryakov et al. 2004], which use domain ontologies for semantically annotating documents and provide a set of general features for managing semantic documents (e.g., documents annotation, storage, indexing and retrieval), being applicable to several domains. These tools provide only general features and do not explore the specific conceptualization provided by the domain ontologies. In order to provide a more effective support to domain-specific tasks, it is useful to explore the ontology elements (concepts, relations and properties) and use them to develop domain-specific functionalities [Falbo et al. 2014].

In this paper, we explore the use of domain ontologies for semantic documentation in Project Management. First, we started by carrying out a systematic literature review (SLR) to analyze initiatives that support project management aspects by using semantic annotation. The use of semantic annotation in the Project Management domain can help project managers to get consolidated information from data stored in different documents and to make decisions based on it. Taking that into account, we aim at extending IMSD to explore specific features to support project management.

This paper is organized as following: Section 2 talks briefly about semantic documentation and project management. Section 3 addresses the performed SLR. Section 4 presents a fragment of the Software Project Management Ontology we developed and discusses its use to extend IMSD. Section 5 concerns related works. Finally, Section 6 presents our final considerations.

2. Semantic Documentation and Project Management

In organizations there is a considerable amount of work done by using desktop tools. Semantic Documentation is a key for tackling the lack of semantics in desktop documents. Semantic documents provide services such as advanced search, reasoning using document metadata, and knowledge management services, like document repositories and document management [Eriksson and Bang 2006].

The problems related to accessing and managing document content clearly occur in the Project Management context, since text documents and spreadsheets are frequently used as instruments for recording and sharing information among project members. In this sense, semantic annotation has potential use in this area.

Project management involves the application of knowledge, skills, tools and techniques to project activities aiming to meet project requirements [PMI 2013]. According to the PMBOK [PMI 2013], there are ten knowledge areas (KAs) related to project management, i.e., there are ten KAs to be managed, namely: Integration, Scope, Stakeholder, Human Resource, Time, Cost, Risk, Quality, Communication, and Procurement.

Project management comprehends three main interactive phases [Pressman 2011]: planning, execution, and monitoring and control. During project planning it is established a plan to the project, including the project scope, allocated human resources, schedule, budget and risks, among others. Execution consists of running the plan, i.e., execute the project following the established plan. In this phase the project results are produced and most of budget and efforts are spent. Monitoring and control aims to compare the plans with the execution, identify problems and present solutions. During this phase, performance indicators can help the project manager to understand the project progress and performance based on the project scope, schedule and budget.

During a project, relevant information regarding planning, progress, monitoring and control is recorded in text documents and spreadsheets (e.g., project management plan and status reports). If information is structured and annotated, computers can help to handle it. Besides, semantic annotation could help store and retrieve the knowledge acquired in a project and reuse it in other projects.

3. Systematic Literature Review

Aiming at identifying and analyzing initiatives involving semantic annotation to support Project Management, we carried out a systematic literature review. According to Kitchenham et al. (2011), systematic literature reviews are secondary studies used to find, critically evaluate and aggregate all relevant research papers on a specific research question or research topic. The methodology is intended to ensure that the literature review is unbiased, rigorous and auditable. The study followed the review process defined by Kitchenham and Charters (2007), which involves three phases: planning, when the research protocol is defined; conducting, when the protocol is executed and data are extracted, analyzed and recorded; and reporting, when the results are recorded and made available. Next, we present the main parts of the protocol used in the study.

3.1 Research Protocol

Research Questions: The main research question is *(RQ1) What are the initiatives involving semantic annotation that support project management aspects?* From this general question, two more specific were defined: *(RQ2) How semantic annotation is addressed?* and *(RQ3) Which are the aspects of project management supported?*

Search String: The search string has two groups of terms joined by the AND operator. The first group aims at capturing studies that deal with semantic annotation and semantic documentation. The second group aims to capture studies related to project management. Within each group, the OR operator was used to allow for alternative terms. The following search string was used: *((("semantic documentation") OR ("semantic annotation") OR ("semantic-document") OR ("semantic document")) AND (("project management") OR ("project planning") OR ("project controlling") OR ("project control") OR ("project monitoring") OR ("project tracking")))*.

Sources: Five digital libraries were searched, namely: Scopus (*www.scopus.com*), Engineering Village (*www.engineeringvillage.com*), ACM (*dl.acm.org*), IEEE Xplore (*ieeexplore.ieee.org*) and ScienceDirect (*www.sciencedirect.com*).

Publications Selection: the object of analysis are articles published in scientific events or journals. Publications selection was done in four steps: the 1st step (S1), *Preliminary*

Selection and Cataloging, consisted in applying the search string by using the digital library search mechanism. Publication language was limited to English, and the search scope was limited to title, abstract and keywords. At the end of this step, publications indexed by more than one digital library were identified and duplications were removed. The 2nd Step (S2), *Selection of Relevant Publications – 1st filter*, involved reading the abstracts of the publications selected in S1 and analyzing them considering the inclusion criterion IC1 - the publication presents some proposal involving semantic annotation that supports aspects related to project management, and two exclusion criteria: EC1 - the publication does not have an abstract; and EC2 - the publication is not a primary study. The 3rd Step (S3), *Selection of Relevant Publications – 2nd filter*, consisted of reading the full text of the publications selected in S2 and analyzing them considering IC1 and other three exclusion criteria: EC3 - the study was published only as an abstract; EC4 - the publication full text is not available; and EC5 - the publication is a copy or an older version of an already selected publication. Finally, in the four step (S4) we performed *Backward Snowballing* [Webster and Watson 2002], investigating if among the references cited in the selected papers, there was some useful to the study.

3.2 Data Synthesis

The systematic review was finished at the beginning of 2015 and considered publications until December 31st 2014. As a result of S1, 39 publications were obtained (21 in Scopus, 13 in Engineering Village, 5 in IEEE). No publication was returned by applying the search string to ACM and ScienceDirect. After duplication removal, 24 publications remained. 21 publications were selected in S2 and 4 in S3. None new paper was selected in S4. The selected papers were published during the last decade, meaning that the research topic is recent. In fact, we expected to find only recent publications, because semantic annotation was applied to semantic documents only in the 2000's. The small number of publications selected shows that, in addition to be recent, the topic has not been much explored. Next, a data synthesis to each research question is presented.

RQ1. What are the initiatives involving semantic annotation that support project management aspects? Four initiatives were found:

- **Semantic Annotation based on Software Knowledge Sharing Space (SKSS)** [Lu et al. 2008]: SKSS is a system that aims to improve knowledge sharing among software development team members. It allows annotating documents produced during projects, creating a network that facilitates accessing and sharing information about the project.

- **Content Management for Inter-Organization Projects (CMIO)** [Nakatsuka and Ishida 2006] : CMIO is a system to manage content of inter-organizational projects. Project content is semantically annotated, and when a project member creates, modifies or manages content in a project, automatic emails are sent to the other project members, communicating explicitly what has changed in the project.

- **Collaboration in Public Policy Making, Implementation and Evaluation (CPPMIE)** [Loukis 2007]: CPPMIE consists of a structured electronic forum in which participants opine about programs, projects, tasks and deliverables related to public policies. A Public Policy Ontology is used for semantically annotating posts, allowing organization, indexing, integration and querying of the posts recorded in the forums.

- **Semex** [Talaš et al. 2010]: Semex is a module of a project management system. It is responsible for semantic annotation of wiki pages. It supports creation, sharing and publication of collaborative content in projects, providing a common environment that allows project team members to access information and contribute to discussions.

RQ2. How semantic annotation is addressed in the initiative?

In this question, we analyzed the semantic annotation approach used in each study, considering aspects such as semantic annotation type, annotated files, ontologies and technologies involved. Regarding semantic annotation type, it is manual when annotations are made by the user. It is automatic when automation components are used to provide suggestions for annotations or make them automatically [Uren et al. 2006].

In SKSS, semantic annotation is used to connect information recorded in different documents. Word, Eclipse, VS.Net and Adobe Reader documents can be annotated. Annotation is manual and based on Project, Annotation and Document domain ontologies. A framework composed of three components is used: the *sensor* component is a plug-in embedded into tools (MS Word, Adobe Reader, Eclipse and Visual Studio) that adds semantic annotations and connects information recorded in different documents; the *service provider* component deals with knowledge publishing, ontology management and query; and the *database* component stores annotation instances, ontologies and documents, and supports version control.

In CMIO, semantic annotation is manual and made by using an application named Project Organizer, which allows for annotating web pages, PDF files and text documents using a Project domain ontology as a basis. CMIO uses e-mail metaphor, i.e., it semantically annotates documents, connects information recorded in different documents, and when document content is created, modified or managed, automatic emails are sent to project members communicating the changes. A RDF database is used to store content, metadata and associations.

CPPMIE annotates web documents and electronic forum pages. The annotation is manual and based on a Public Policy domain ontology. A structured electronic forum based on the ontology is used to record posts about public policies projects and programs. Information semantically annotated in posts is retrieved and an XML file containing relevant information is produced.

Semex annotates wiki pages, allowing for browsing pages containing project content and selecting information related to the projects (e.g., projects that share a certain human resource). Semantic annotation is manual and uses a Project Management and Presentation domain ontology as a basis. Semex uses RDF triple to annotate wiki pages and RDFLib library (*www.rdflib.net*) to work with RDF.

RQ3. Which are the aspects of project management supported by the initiative?

Aspects related to four KA are supported by the initiatives: Scope, Integration, Communication and Stakeholder Management.

Communication Management KA covers communication planning (definition of what information should be available; how, when and where it should be recorded; who is responsible for recording it; and who can access it), management (communication plan execution) and controlling (comparison between planned and executed, and

corrective actions execution). Three proposals support this KA, mainly in aspects related to communication management, which occurs during the project execution phase. In SKSS, semantic annotation helps information recording and sharing. For instance, documents produced during the project can be annotated and related one to others in a knowledge network. As a result, when a document is accessed by a project member, she also gets its related documents. In Semex, a common knowledge base is shared between projects and supports information sharing. Semantic annotation allows for browsing pages containing project content and selecting information related to the projects (e.g., projects that share a certain human resource). CMIO supports project content creation, modification and management, and sends automatic emails to project members communicating the changes made. By doing this, CMIO also supports aspects related to Integration Management that includes, among others, integrated change control, consisting of recording the project changes, their reasons, and performing the necessary actions in an integrated way.

CPPMIE supports Scope and Stakeholder Management aspects. Scope Management concerns the definition of the work to be done in the project, while Stakeholder Management involves identifying and managing project stakeholders, their expectations and involvement. The CPPMIE forum is used to define the public policies and requirements to be addressed in projects, i.e., the project scope. Moreover, the forum helps to interact with stakeholders, encouraging the appropriate involvement of them in project activities.

3.3 Discussions

By analyzing the selected papers, we noticed that, except by Semex, the proposals were not conceived aiming to support project management. Thus, although the proposals support aspects related to project management, this is not their main concern.

Regarding the semantic annotation approach adopted, all proposals use domain ontologies as a basis for annotating documents or web pages. Spreadsheets are not annotated in any proposal. Also, all proposals adopt manual annotation. According to Uren et al. (2006), automation is a desirable requirement in semantic annotation proposals. Manual annotation is an additional burden, because human annotators are prone to error and non-trivial annotations usually require domain expertise. However, there are research challenges in this direction, related to the extraction of relations for semantic annotation.

As for the project management aspects addressed, the proposals support some ones related to Scope, Integration, Communication and Stakeholder Management. Since Communication Management is related to information recording and sharing, and semantic annotation supports them, it was expected that Communication was among the main supported areas. The other knowledge areas that are supported by the proposals usually produce documents as results of their activities (e.g., requirements document produced in Scope Management). Time and Cost Management, which are important areas in project management, are not supported by any proposal. Semantic annotation could help relate and sequence the project activities and control the schedule. Besides, it could support cost and quality control, for example, by establishing relationships between costs and activities, and between changes and deliverables. However, these

KAs are typically well supported by project management systems (e.g., MSProject). This can be one of the reasons why these areas have not been target of semantic annotation initiatives. Besides, the use of semantic annotation in project management is very recent. Thus, there are still many aspects to be explored.

As limitations of this systematic review, we highlight the small number of selected publications. Although five digital libraries have been used, only four publications were identified and only one of them is truly devoted to the project management domain. This fact shows that the research topic is recent and has not been much explored. Since documents are still an important instrument to record and share information regarding projects, we believe that the use of semantic annotation on project management is a relevant topic, and there are opportunities of research in this area.

4. Using Semantic Annotation to support Project Management

In order to explore the use of semantic annotation in the project management context, we extended the Infrastructure for Managing Semantic Documents (IMSD) [Arantes and Falbo 2010]. IMSD provides: *(i)* a way to semantically annotate document templates; *(ii)* a mechanism for controlling versions of semantic content extracted from semantic document versions, and therefore providing a way for tracking the evolution of the data embedded inside a semantic document; and *(iii)* data visibility to end-users allowing searches and data change notification subscription to aid developers to get an up-to-date information about something they are interested in.

IMSD supports the use of templates in text format. Since spreadsheets are very useful for recording data regarding projects (e.g., schedules and budges), we decided to extend IMSD to work with spreadsheets, expanding the scope of files used as data sources. Moreover, in order to annotate document and spreadsheet templates with metadata related to software project management, we developed the Software Project Management Ontology. Thus, we explored its conceptualization in domain-specific features to support project management activities.

4.1 The Software Project Management Ontology

The Software Project Management Ontology (SPMOnt) was developed based on the Software Process Ontology Pattern Language (SP-OPL) proposed in [Falbo et al. 2013]. SPMOnt includes concepts, relations and properties related to scope, time and costs planning and execution. Regarding costs, currently, only costs associated with human resources are considered. Figure 1 shows a fragment of SPMOnt with some of the concepts related to time and cost planning and execution. SPMOnt is represented by using OntoUML, a UML profile that enables modelers to make finer-grained modeling distinctions between different types of classes and relations according to ontological distinctions put forth by the Unified Foundational Ontology [Guizzardi 2005].

There are two types of processes defined to a **Project**: **General Project Process** and **Specific Project Process**. The first one is the global process defined to the Project. It is composed by specific process, allowing defining **sub-processes**. Specific Project Processes are composed by **Project Activities**, which can be **Simple Project Activities** or **Composite Project Activities**. Once a general project process is defined to a project, it is possible to plan duration, start and end dates, and cost of the process, their sub-

processes and activities. The definition of duration, dates and cost to a Project Process gives rise, respectively, to **Process with Planned Duration**, **Scheduled Process** and **Process with Planned Cost**. Similarly, the planning of duration, dates and cost of a Project Activity gives rise to **Activity with Planned Duration**, **Scheduled Activity** and **Activity with Planned Cost**.

A **Human Resource Allocation** is the assignment of a Scheduled Activity to a **Human Resource** to perform a **Human Role**. The cost of a Human Resource Allocation is based on the cost of the allocated Human Resource, which is established in the **Employment** of that Human Resource.

A Project Activity can cause **Activity Occurrences**, which can be **Simple Activity Occurrences** or **Composite Activity Occurrences**. **Human Resource Participation** refers to the participation of a Human Resource in an Activity Occurrence.

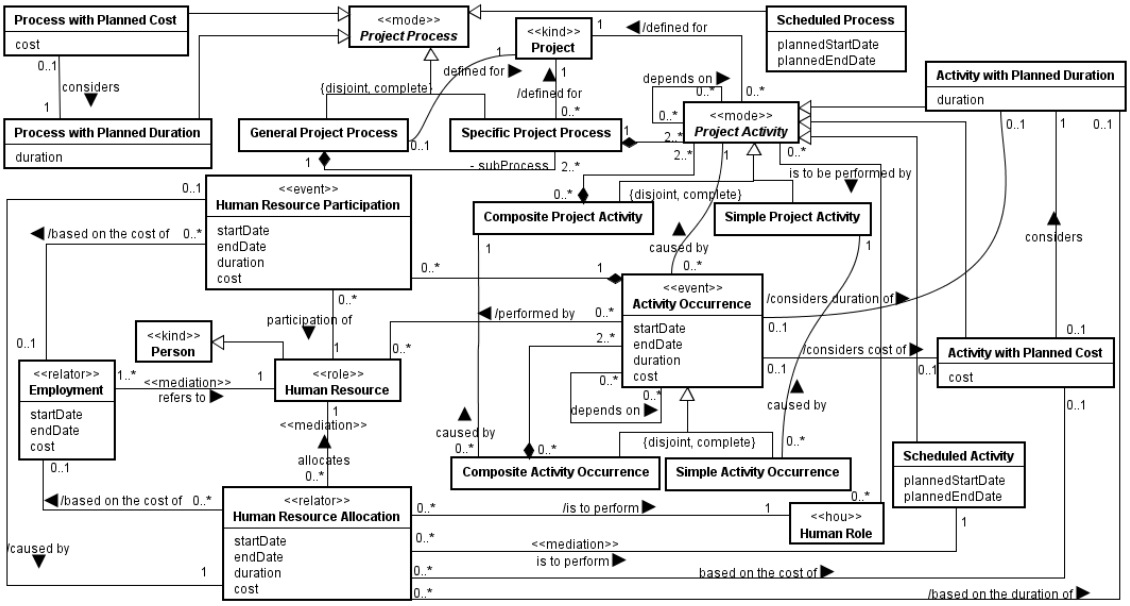

Figure 1 – A fragment of the Software Project Management Ontology

4.2 Supporting Project Management with Semantic Annotations in Spreadsheets

In order to explore the use of semantic annotation to support project management aspects, we first extended IMSD to work with spreadsheets and then we used SPMOnt as a basis to annotate spreadsheet templates related to the project management domain. The annotations are added into the templates that, when instantiated, give rise to semantic spreadsheets. Thus, once annotated the templates, the spreadsheets produced using them are also annotated and can be used as data sources to IMSD. Spreadsheet templates were developed using the Open Document Format [Oasis 2015], since it is an open format, with great span. Specialized annotations for cells were produced using Open Document Spreadsheet (ODS) in LibreOffice Calc.

For spreadsheets annotation, the syntax and instructions for annotating text fragments provided by IMSD are used to capture the cell content. Instructions can be used to create instances, relations and properties based on the ontology. The syntax of the instance creation instruction is *instance (arg ,concept, accessVariable)*. This instruction

creates the instance *arg* of the *concept* of SPMOnt. The SPMOnt was implemented in OWL and its URL is also informed in the *concept* field. The instruction result is a reference to the created instance and it is set on the *accessVariable* for later use. The syntax to create a relation is *property (arg1, prop, arg2)*. This instruction establishes a relation *prop* between the instances *arg1* and *arg2*. This instruction is also used to create properties and, in this case, it means that the value *arg2* is set as the property *prop* of the instance *arg1*.

For annotating templates and allowing the capture of the spreadsheets content by IMSD, in the LibreOffice Calc, *Custom Properties* option is used to annotations recording and *Styles and Formatting* option is used to allow for application of annotations to cells. The first thing to do when creating a semantic template is to create a custom property named *Semantic Document* and set its value to *True*. This way, IMSD can identify that the spreadsheet is a semantic document and searches for semantic annotations. Each annotation must be recorded in a new custom property whose value is the annotation instruction. For each annotation, a formatting style must be created and it must be related to the custom property in which the annotation is recorded. Thus, when a formatting style is applied to a cell, the cell is annotated according to the annotation instruction recorded in the corresponding custom property.

Three templates related to project management were developed and annotated: *WBS*, which is a text document that describes the project deliverables and work packages; *Project Status Report* (PSR), which is a spreadsheet that contains information regarding project planning and execution; and *Human Resources Costs* (HRC), which is a spreadsheet that provides information regarding the costs of human resources allocated to the project. Figure 2 shows the template of the *Project Status Report*, which contains information about project activities, dependencies, human resources allocated and participants, WBS items related, and planned and executed dates and duration. As examples, the annotations related to cells of Human Resource and Duration columns are shown. The first part of the human resource annotation creates instances of the *Human Resource* concept and stores in *hr* variable. The second part establishes the relationship *allocates* between instances of Human Resource and an instance of Activity, like in SPMOnt, in which the relation *allocates* connects a human resource to an activity, meaning that the human resource is allocated to perform the activity. The *break* tag means that one or various human resources can be related to one activity and they are separated by comma. In duration annotation, the tag *completeText* indicates that the instruction refers to the complete text stored in the cell. The instruction means that the cell content will be set as the property *Planned Duration* of an instance of Activity.

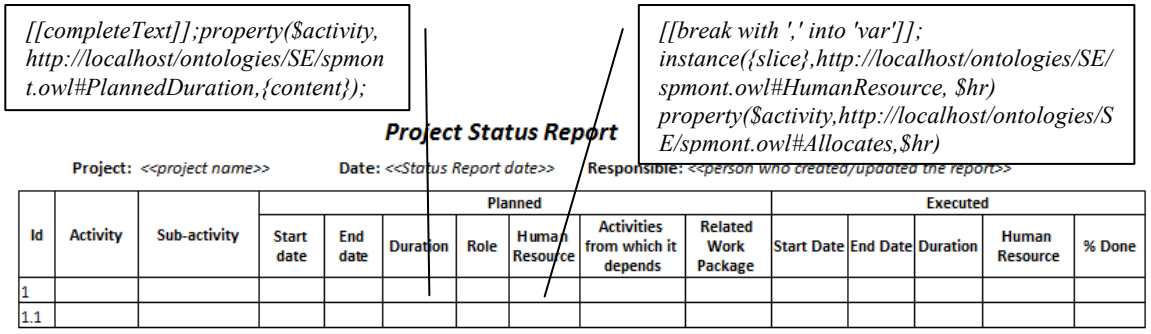

Figure 2 – Project Status Report template

The spreadsheets produced using the annotated templates are submitted to IMSD, which extracts data from them and stores in OWL files, allowing searching and retrieval. IMSD also performs version control of the spreadsheets and notifies users about changes. Annotation, indexing, storing, retrieval, version control and changes notification are general functionalities, which can be applied to any domain.

We argue that project management aspects can be better supported by exploring the conceptualization provided by the domain ontology. In this sense, some domain-specific functionalities were identified from the SPMOnt concepts, relations and properties, and have been implemented to extend IMSD: *(i)* the dependency relation between activities and between activities and WBS items (not shown in Figure 1) can be used to extract and relate data recorded in Project Status Reports and WBS document and represent them in dependency matrices that are useful to analyze the impact of changes in the project; *(ii)* the relation between activities and project cost with the human resource allocations cost can be explored to, based on activity duration, human resources allocations and human resources costs, define the project budged; *(iii)* relationships between activities with planned duration/cost and the real duration/cost of the activity occurrences caused by them can be explored to track planned and executed values, determine their adherence, and also calculate Earned Value Analysis indicators and estimates about the project conclusion, helping project managers to understand the project progress, monitor it and make adjustments when necessary; and *(iv)* indicators calculated to several projects can be represented in graphics allowing project managers to have a global view of the projects and make comparisons among them.

5. Related Works

As discussed in Section 3, there are some initiatives involving semantic annotation that support project management aspects. There are some similarities between our work and the proposals found in the systematic review. However, there are also differences.

As for similarities, like IMSD, all proposals use domain ontologies as a basis to annotations and provide general features for managing semantic content (annotation, storage, indexing and retrieving). Based on the semantic content, SKSS [Lu et al. 2008] creates a knowledge network of documents. Similarly, IMSD uses semantic content and creates graphs in which information recorded in documents are related one to another. CMIO [Nakatsuka and Ishida 2006] and IMSD send automatic emails notifying users about modifications on semantic documents.

The main differences between our proposal and the ones found in the SLR concern the types of annotated files and the project management knowledge areas supported. Regarding types of files, the proposals annotate web pages, electronic forums, pdf and text documents. IMSD also annotates text documents, but it is the only one to annotate spreadsheets.

As for the knowledge areas supported, as discussed in Section 3, the proposals support aspects related to Scope, Integration, Communication and Stakeholder Management. IMSD, in turn, deals with aspects related to Scope, Time and Costs Management. Thus, IMSD differs from the cited proposals mainly due to the features to support project management activities, obtained by exploring the SPMOnt conceptualization in functionalities that help managers to plan, monitor and control

projects. Although the proposals support some project management aspects, the domain ontologies used do not address aspects that allow for comparing project planning and execution. Also, none proposal provides indicators or estimates to help project managers to monitor projects. Summarizing, by exploring the SPMOnt conceptualization, domain-specific features are provided by IMSD, better supporting project management activities.

6. Final Considerations

In this paper we discussed the use of semantic annotation in project management. The results of a systematic literature review that investigated initiatives that support project management aspects by using semantic annotation were presented. We also discussed an extension of the IMSD [Arantes and Falbo 2010] that enables it to semantically annotate spreadsheets with concepts, relations and properties of the Software Project Management Ontology to provide features supporting project planning and tracking.

At this moment, we are concluding the implementation of the ISMD domain-specific functionalities. As future work, we plan to conduct experiments to evaluate the extension of IMSD in the project management domain. Moreover, we intend to integrate project management tools (such as MS-Project) with documents and spreadsheets semantically annotated by IMSD. By doing this, organizations that use these tools can also benefit from IMSD functionalities. Finally, we intend to improve cost management features by considering costs relate to software, hardware and other cost elements that have not been currently considered.

Acknowledgment

This research is funded by the Brazilian Research Funding Agency CNPq (Processes 485368/2013-7 and 461777/2014-2).

References

ARANTES, L. O. and FALBO, R. A. (2010) "An infrastructure for managing semantic documents", In: Joint 5th International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE) - International Workshop on Metamodels, Ontologies and Semantic Technologies (MOST), p. 235-244.

BRUGGEMANN, B. M., HOLZ, K.-P. and MOLKENTHIN, F. (2000) "Semantic documentation in engineering", Eighth International Conference on Computing in Civil and Building Engineering, California, USA, p. 828-835.

ERIKSSON, H. (2007) "The semantic-document approach to combining documents and ontologies", International Journal of Human-Computer Studies, v. 65, n. 7.

ERIKSSON, H. and BANG, M. (2006) "Towards document repositories based on semantic documents", Sixth International Conference on Knowledge Management and Knowledge Technologies (I-KNOW), Graz, Austria, p. 313-320.

FALBO, R. A. et al. (2013) "Organizing Ontology Design Patterns as Ontology Pattern Languages", 10th European Semantic Web Conference – ESWC 2013, France, p. 61-75.

FALBO, R. A., BRAGA, C. E. C. and MACHADO, B. N. (2014) "Semantic Documentation in Requirements Engineering", In: 17th Workshop on Requirements Engineering (WER 2014), Pucón - Chile,

GUIZZARDI, G. (2005) "Ontological Foundations for Structural Conceptual Models", University of Twente, The Netherlands.

KIRYAKOV, A., POPOV, B. and TERZIEV, I. (2004) "Semantic annotation, indexing, and retrieval", Web Semantics: Science, Services and Agents on the World Wide Web, v. 2, p. 49-79.

KITCHENHAM, B. and CHARTERS, S. (2007) "Guidelines for performing systematic literature reviews in software engineering", (EBSE-2007-01)

KITCHENHAM, B. A., BUDGEN, D. and BRERETON, O. P. (2011) "Using Mapping Studies as the Basis for Further Research - A Participant-Observer Case Study", Information & Software Technology, v. 53, n. 6, p. 638-651.

LOUKIS, E. N. (2007) "An ontology for G2G collaboration in public policy making, implementation and evaluation", Artificial Intelligence and Law, v. 15, n. 1, p. 19-48.

LU, Q., CHEN, M. and WANG, Z. (2008) "A semantic annotation based software knowledge sharing space", In: IFIP International Conference on Network and Parallel Computing (NPC), China, p. 504 - 509.

NAKATSUKA, K. and ISHIDA, T. (2006) "Content management for inter-organizational projects using e-mail metaphor", International Symposium on Applications and the Internet (SAINT), Phoenix, Arizona, USA, p. 202-205.

OASIS. Open Document Format for Office Applications. Visited in: July, 9th 2015, https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office.

PMI (2013), A guide to the Project Management Body of Knowledge (PMBoK), Project Management Institute, 5.

PRESSMAN, R. S. (2011), Engenharia de Software, McGraw Hill, 7th edition.

SICILIA, M. (2006) "Metadata, semantics and ontology: providing meaning to information resources", International Journal of Metadata, Semantics and Ontologies, v. 1, n. 1, p. 83-86.

TALAŠ, J., GREGAR, T. and PITNER, T. (2010) "Semantically enriched tools for the knowledge society: case of project management and presentation", Third World Summit on the Knowledge Society, Greece, p. 322-328

UREN, V. et al. (2006) "Semantic annotation for knowledge management: requirements and a survey of the state of the art", Journal of Web Semantics: Science, Services and Agents on the World Wide Web 4, p. 14-28.

VILLALOBOS, J., SANABRIA, S. and CACERES, R. (2011) "Activity scheduling through gantt charts in an ms excel spreadsheet", Revista Facultad de Ingenieria, n. 61, p. 132-145.

WEBSTER, J. and WATSON, R. T. (2002) "Analyzing the Past to Prepare for the Future: Writing a Literature Review", MIS Quarterly, v. 26, n. 2, p. 13-23.

An Ontology for Collaborative Tasks in Multi-agent Systems

Daniela Schmidt [1], Rafael H. Bordini [1], Felipe Meneguzzi [1], Renata Vieira [1]

[1]Postgraduate Program in Computer Science
School of Informatics (FACIN)
Pontifical Catholic University of Rio Grande do Sul - PUCRS
daniela.schmidt@acad.pucrs.br
{rafael.bordini,felipe.meneguzzi,renata.vieira}@pucrs.br

Abstract. *This paper proposes an ontology for task representation and inference. The ontology was developed to support reasoning about tasks, such as task recognition and relocation. Our proposal is formalized in OWL (Web Ontology Language) and SWRL (Semantic Web Rule Language). We show one scenario to exemplify reasoning situations based on the axioms and rules of our ontology. This knowledge-level representation of tasks can be explored to support reasoning about activities for groups of people. The knowledge asserted and inferred in the ontology is useful in multi-agent systems to enhance agent coordination and collaboration through reasoning over tasks. An evaluation of the proposed ontology is presented.*

1. Introduction

Ontology and agent-based technologies have received significant attention, but little focus has been given in their integrated use [Hadzic et al. 2009a]. Ontologies allow the sharing of common knowledge among people and software agents. For multi-agent systems (MAS) development, ontologies are key for the common understanding and reuse of domain knowledge.

Our work aims to integrate an ontology that represents collaborative tasks with a multi-agent framework to enable queries and inferences about shared tasks. It provides knowledge about tasks for the execution of plan recognition, and for the negotiation and relocation of tasks. In this paper, we describe our task ontology in detail, as well as its integration in a multi-agent system. Based on the proposed ontology, we present an application in *health care* which consists of a family group that takes care of an elderly person. The elderly needs constant monitoring to perform his daily tasks, so it is necessary that the group collaborates in the distribution of tasks related to his care.

The paper is organized as follow: Section 2 introduces main aspects of ontologies, and technological alternatives for their representation such as OWL and SWRL. Section 3 describes related work. Next, in Section 4 we present our proposed ontology. Section 5 presents the ontology evaluation. Section 6 shows some final remarks and future directions to extend our proposed ontology.

2. Ontology

Ontology was originally the philosophical study of reality to define which things exists and what we can say about them. In computer science, ontology is defined as an "explicit specification of a conceptualization" [Gruber 1993]. A conceptualization stands for an

abstract model of some world aspect that specifies properties of important concepts and relationships. Therefore, ontologies are knowledge representation structures composed of concepts, properties, individuals, relationships and axioms. A *concept* (or class) is a collection of objects that share specific restrictions, similarities or common properties. A *property* expresses relationships between concepts. An *individual* (instance, object, or fact) represents an element of a concept; a *relationship* instantiates a property to relate two individuals; and an *axiom* (or rule) imposes constraints on the values of concepts or individuals normally using logic languages (which can be used to check ontological consistency or to infer new knowledge).

Nowadays there are prominent ontology languages, such as OWL (Web Ontology Language) [Bechhofer et al. 2004], which is a semantic web standard formalism to explicitly represent the meaning of terms and the relationships between those terms. OWL is a language for processing web information that became a W3C recommendation in 2004 [Bechhofer et al. 2004]. Ontologies empowers the execution of semantic reasoners, such as Pellet [Sirin et al. 2007]. Semantic reasoners provide the functionalities of *consistency checking*, *concept satisfiability* and *classification* [Sirin et al. 2007]. In other words, reasoners infer logical consequences from a set of axioms, which in the current technology can be done, for example, through the application of the rules coded in SWRL (Semantic Web Rule Language) [Horrocks et al. 2004].

Ontologies and rules are established paradigms in knowledge modeling that can be used together. SWRL is a rule extension of OWL that adheres to the open-world paradigm [Horrocks et al. 2004]. SWRL adds to the OWL's expressiveness by allowing the modeling of certain axioms which lie outside the capability of OWL; including an abstract syntax for Horn-like rules in ontologies. The rules are defined as an implication between an antecedent (body) and a consequent (head). The intended meaning can be read as: whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold (be true). A SWRL rule [Horrocks et al. 2004] has the form of $antecedent \Rightarrow consequent$ where both antecedent and consequent are conjunctions of atoms written as $a1 \wedge ... \wedge an$. Variables are prefixed with a question mark (*e.g.*, `?x`), and an atom may be unary (*e.g.*, a class expression) or binary (*e.g.*, an object property), and arguments in atoms can be individuals or data values. In this syntax, a rule asserting that the composition of parent and brother properties implies the uncle property can be written as [Horrocks et al. 2004]: $parent(?x, ?y) \wedge brother(?y, ?z) \Rightarrow uncle(?x, ?z)$.

This section provided a background on ontology technologies employed in this work. Our proposal is formalized in OWL and SWRL, and applies the semantic reasoner Pellet [Sirin et al. 2007] for executing inferences over the modeled concepts, properties, individuals, axioms and rules. Next, we present previous approaches to task ontologies.

3. Related Work

This section describes previous work that use ontologies for activity representation and recognition. Activity modeling in ontologies consists in defining formal semantics for human tasks by means of the operators in ontological languages. Then, ontological reasoning can be used to recognize that the user is performing a certain activity starting from some facts (*e.g.*, sensor data, location of persons and objects, properties of actors

involved) [Riboni and Bettini 2011b].

Smith *et al.* [Smith et al. 2011] propose an ontology containing rules to reason about task characteristics to enable an effective coordination of activities. To implement such mechanism, agents need to reason and communicate about activities, resources and their properties. The semantics for the coordination mechanisms are given by rules coded in an ontology. The idea is to enable agents to reason about the relationships of their activities with the activities of other agents [Smith et al. 2011].

The OWL-T ontology [Tran and Tsuji 2007] was designed to allow the formal and semantic specification of tasks using a high-level knowledge abstraction. The *Task* class is the central concept in OWL-T, and can be hierarchically decomposed into simple or complex tasks. According to Tran and Tsuji [Tran and Tsuji 2007], an example of complex task is planning a trip, that requires the sub-tasks of book flight, book hotel and rent car.

Riboni and Bettini [Riboni and Bettini 2011a] propose an ontology of activities for task representation that combines ontological reasoning with statistical inference to enable activity recognition. Their solution uses statistical inference on raw data retrieved from body-worn sensors (*e.g.*, accelerometers) to predict the most probable activities. Then, symbolic reasoning refines the results of statistical inference by selecting the set of possible activities performed by a user based on the context. The ActivO (ontology for activity recognition) contains a set of activities, as well as context data that can be useful to recognize them. The set of activities and context data defined in ActivO is non-exhaustive, but it is claimed that it can be used to model many pervasive computing scenarios.

The work on [Chen et al. 2012] introduces a knowledge-driven approach to activity recognition and inferencing based on multi-sensor data streams in smart homes. The ontology represents the correlated domains of smart homes contexts and Activities of Daily Living (ADL). For example, the task of brushing teeth normally takes place in the bathroom twice a day and usually involves using toothpaste, toothbrush and water, which is the context for this activity. Contextual information is obtained by sensors that are linked to physical and conceptual entities such as objects, locations, and states. In addition to the ontology of contexts, an ontology for activity recognition was created to model an activity hierarchy in which each class denotes an ADL type. ADL ontologies can be viewed as activity models that establish links between activities and contextual information.

Garcia *et al.* [Garcia et al. 2013] propose one approach based on ontologies to solve problems related with resource sharing in pervasive environments. The ontological model is composed by a set of ontologies that represent the elements involved in a collaborative environment. Ontologies refers to types of managed resources (human, physical and virtual) and other characteristics such as environment and organizational aspects. This set of ontologies is part of RAMS architecture (Resource Availability Management Service). According to the authors, a set of ontologies is better than one unique ontology and these proposed ontologies can be extend by adding new concepts.

Bae [Bae 2014] presents and approach to activities of daily living (ADL) recognition called RADL (Recognizing Activities of Daily Living). RADL is a system that

Table 1. Comparing the domains of related ontologies

Ontology	Scope	Main Concepts
Coordination ontology [Smith et al. 2011]	Tasks and their positive/negative relationships to enable agent coordination.	**Activity**, interdependence types, agent, resource, operational relationship, and coordination rule.
OWL-T [Tran and Tsuji 2007]	Business processes using high-level abstraction.	**Task** (divided into simple and composite), input, output, pre/post conditions, preferences, effects.
ActivO [Riboni and Bettini 2011a]	Concepts to help in activity recognition, specially in the health care field.	**Activity**, artifact, communicationRoute, **person**, symbolicLocation, timeExtend.
SH and ADL [Chen et al. 2012]	Daily-living activities in the smart home domain.	**Activity**, actor, **location**, resource, environment, entities, duration, goal, effects, conditions, time.
RAMS [Garcia et al. 2013]	Represent actions that users can execute about resources.	Process, **activity**, group, role, calendar, time interval, human resource.
RADL [Bae 2014]	Represent activities of daily living related to elderly person in a smart home environment.	**Person**, **activity**, **location**, device, device status, sensor, sensor status, service, daily life service, message service, safety service.

detects and monitors ADL's standards for smart homes equipped with sensors. RADL is exemplified in a smart home scenario where one elderly person lives alone. The ontology proposed by the author is able to reason about ADL's standards and provide semantic discovery of locations, devices, activities and other relevant information. The ontology is divided in three parts. The first represents concepts about daily life services like: air conditioner on or off and open or closed window for instance. The second represents safety services like: fire alarm activated. And the third part represent messages services like: sleeping message, wake up message and so on are described.

As we see, semantic representations of tasks through ontologies are starting to appear as promising research directions. These representations enable agents to reason about tasks, for example, to implement activity recognition approaches. In Table 1 we compare the related work presented in Section 3, highlighting the main concepts of each ontology. In our proposed ontology, we reused the most common concepts in the domain (as highlighted in bold in Table 1). We then included new aspects to those found in previous works, since our focus was on the representation of tasks performed collaboratively (to be presented in the next section). No previous work was found with such orientation.

4. Proposed Task Ontology

This section presents our proposed ontology, the integration of our ontology in the multi-agent system and its application. The goal is to represent the knowledge of where and when the tasks might occur, who is responsible for them and what are the tasks particularities. Based on this ontology representation, we may use logical rules and apply semantic reasoners to infer new knowledge about tasks that may be useful for agent programmers willing to implement task reasoning mechanisms, such as techniques of task recognition, task negotiation and task relocation.

The *Task* is the main concept in the ontology; it represents an activity that is executed by one or more people. We can also say that the execution of a *Task* may happen in a particular location and time, and normally involves an object. Therefore, the main and most generic concepts of the proposed task ontology are: *Task*, *Person*, *Location*, *Object*, *TimeInterval* and *TaskPurpose* (see Figure 1 a)).

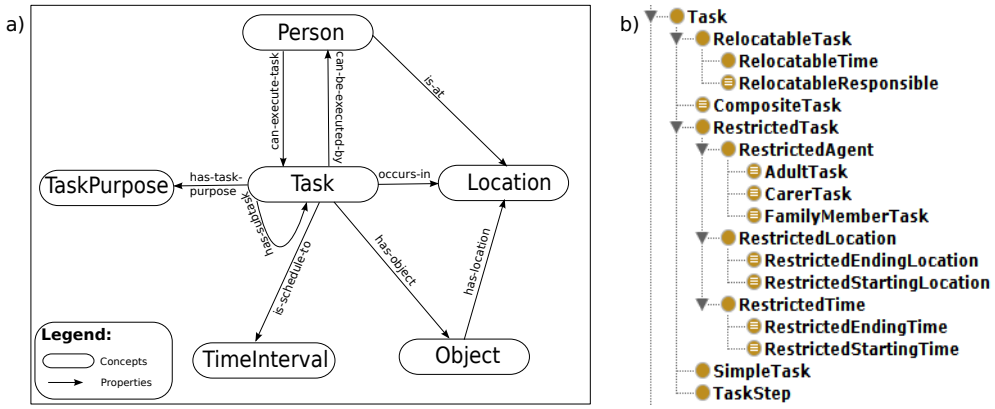

Figure 1. a) Task ontology main concepts. b) Taxonomy of task concepts

Collaborative tasks may have restrictions as to who can execute them, when and where they occur. Then, to address these issues our ontology specialized the task concept (according to Figure 1 b)). In our ontology, the concepts were defined based on restrictions and other logical characteristics related to collaborative tasks. For example, the *CompositeTask* concept is equivalent to a task that has sub-task. This concept definition used the existential quantifier, as follows:

$$CompositeTask \equiv \exists has\text{-}subtask.Task$$

The *RestrictedTask* concept is subdivided in three kinds of restrictions that are presented bellow:

RestrictedAgent: in this case, the concepts can be used to define features that agents or people may to perform certain tasks. In our ontology there are three concepts to specify restrictions regarding agents. They are: *(i)* the concept *AdultTask* restricts the tasks that may be performed only by adults, like driving a car, for instance. *(ii)* the concept *CarerTask* restricts the tasks that may be performed only by carers, *(iii)* the concept *FamilyMemberTask* that represents tasks that can only be performed by family members. The setting of restrictions on concepts specified above is as follows:

$$AdultTask \equiv \forall \textit{can-be-executed-by}.Adult$$
$$CarerTask \equiv \forall \textit{can-be-executed-by}.Carer$$
$$FamilyMemberTask \equiv \forall \textit{can-be-executed-by}.FamilyMember$$

Note that the restrictions regarding agent may vary according to the application, in the case of our application carer and family members are types of agents, but in other applications these types may differ. More details on the application is described in the sequence.

RestrictedLocation: a task can be classified according to the location where it occurs. To represent these restrictions the concepts *RestrictedStartingLocation* and *RestrictedEndingLocation* were included. One task instance belongs to any of these concepts if it has the property *possible-starting-location* or *possible-ending-location* respectively. This definition is specified as follows:

$$RestrictedStartingLocation \equiv \exists \textit{possible-starting-location}.Location$$
$$RestrictedEndingLocation \equiv \exists \textit{possible-ending-location}.Location$$

RestrictedTime: similar to the location constraints, a task can have time restrictions. The concepts *RestrictedStartingTime* and *RestrictedEndingTime* are used when a task has restrictions regarding the start or end time of it execution (a physiotherapy session, for instance). A task instance is classified as *RestrictedTime* if it has any restrictions through the properties *has-beginning* or *has-end* respectively.

In addition to the specializations and restrictions, our task ontology allows agents to negotiate about relocation of tasks. To provide this information, the concept *Task* has the sub-concept *RelocatableTask* whose function is to describe the possibilities of relocation in terms of time and responsible for the execution of a task.

RelocatableTask: this concept describes the relocation task possibilities. It is divided in two sub-concepts called *RelocatableResponsible* and *RelocatableTime*. The first refers to a task instance which has the property *can-be-relocated-to*. Rules in SWRL were created to define who is able to perform each task. For instance, some tasks can be executed only by adults. Already, the temporal relocation can occurs when one task instance is not *RestrictedTime*. The following we define the concept *RelocatableResponsible* and the rules in SWRL that allow inferences instances of people for that a task can be relocated.

$$RelocatableResponsible \equiv \exists \textit{can-be-relocated-to}.Person$$
$$Task(?x) \wedge Person(?y) \Rightarrow \textit{can-be-relocated-to}(?x, ?y)$$
$$AdultTask(?x) \wedge Adult(?y) \Rightarrow \textit{can-be-relocated-to}(?x, ?y)$$

Furthermore, a *Task* may contain restrictions based on locations where it can happen, what objects are related, and so on (see Figure 1). These aspects are addressed below.

The *Location* concept represents physical places where *Task* instances happens.

Table 2. Domain and range of task ontology properties

Domain	Object Property	Range
Task	has-subtask	Task
Task	is-part-of	Task
Task	can-be-execute-by	Person
Task	can-be-relocated-to	Person
Task	has-feature	Feature
Task	has-object	Object
Task	has-task-purpose	TaskPurpose
Task	occurs-in	Location
Task	is-schedule-to	TimeInterval
Person	can-execute-task	Task
Person	is-at	Location
Object	has-location	Location
Object	is-used-for	Task
Patient	has-carer	CarerGroup
CarerGroup	carer-of	Patient

Location has two sub-concepts to differentiate between internal and external locations. The relationship of the task concept and location is given by the property *occurs-in*.

The *Person* concept represents the group of people.The relationship between Task and Person occurs through the property *can-be-executed-by*. In order to specialize the ontology according to the application, the *person* concept has two sub-concepts called *CarerGroup* and *Patient*. The first is divided into *Carer* and *FamilyMember*. Instances of carer group are responsible to take care of patient instances. To provide this relationship, we create the object property *has-carer* (see following).

$$Patient \equiv \exists has\text{-}carer.CarerGroup$$

The *Object* concept represents the objects involved in the task execution. The *TaskPurpose* concept represents the specialization of tasks (*e.g. entertainment, hygiene, etc.*). The relationship between this concept and the Task concept occurs through the property *has-task-purpose*.

The *TimeInterval* concept includes information about temporal restrictions. This concept is related to task concept through of the property *is-schedule-to*. *TimeInterval* has three sub-concepts called *ClosedInterval*, *LeftClosedInterval* and *RightClosedInterval*. A task instance is classified with restricted time according to follow rules.

$$ClosedInterval \equiv \exists has\text{-}beginning.string \wedge \exists has\text{-}end.string$$

$$LeftClosedInterval \equiv \exists has\text{-}beginning.string$$

$$RightClosedInterval \equiv \exists has\text{-}end.string$$

The relationship between the concepts occurs through of properties. In Table 2 we

present the main properties of our proposed ontology.

4.1. Ontology Integration in the Multi-agent System

The task ontology is part of a multi-agent framework and allows queries and inferences about tasks in a multi-agent environment. It provides knowledge to plan recognitions and task negotiation and relocation. Figure 2 shows a view of the application of our ontology in the multi-agent framework (more details about the framework and the artifact which allows agents to interact with ontologies can be found in our other papers [Freitas et al. 2015], [Panisson et al. 2015]).

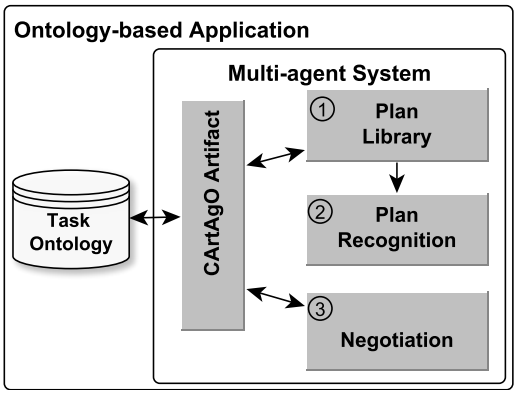

Figure 2. Application of task ontology in the multi-agent framework

The *Plan Library (1)* can be created by instances and restrictions of tasks modeling in the ontology. Consider a plan called *prepare-meal*. In the ontology, there is a instance of *CompositeTask* called prepare-meal. This instance is classified as a top-level plan in the plan library and is decomposed with the sub-tasks instances like: prepare-breakfast, prepare-lunch and so on. The hierarchy between the top-level plan and the sub tasks occurs through the has-subtask property that allows differentiate between sequence or decomposition. In Figure 3 we show a plan makes by ontology instances.

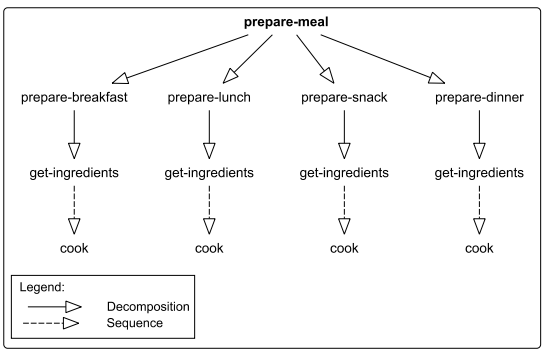

Figure 3. Plan *prepare-meal*

The *Plan Recognition (2)* module is responsible to recognize the agent plans. For this, we used a plan recognizer developed by Avrahami-Zilberbrand and Kaminka [Avrahami-Zilberbrand and Kaminka 2005]. The plans are based on the structure specified in the plan library that was generated from the ontology. In this context, ontology's

role is to provide subsidies for the construction of plans and a set of features that help the plan recognizer to identify what task is running, if it will fail or if the plan needs to start a process of negotiation.

The *Negotiation (3)* module performs queries in the ontology to verify two types of information. *(i)* when agents need to know if a task is relocatable temporally and *(ii)* when agents need to know if a task can be relocated to another member of the group. In the first case, it is checked if the task instance belongs to the concept *RelocatableTime*, if the answer is positive, the task can be relocated temporarily. In the second case, the agent asks the ontology if one task instance has the property *can-be-relocated-to*. This property relates a task instance that can be relocated to other group members who are able to execute it. In this case, the application proceeds to relocation between the members related to the task. The existence of this property takes into account constraints such as tasks that can be performed only by adults, for example.

4.2. Task Ontology Application

According to Bae [Bae 2014], recognition and monitoring of daily activities can provide important opportunities for applications whose focus is the care of the elderly. Moreover, according to the author, the correct way to represent knowledge in household, including behavioral rules systems, is through concepts and information modeled in ontologies. In the other words, ontologies provide readable and understandable knowledge for machine and perform an important role in knowledge representation, sharing and data management, information retrieval, among others. Computational agent-based systems are used to support distributed computing, dynamic information retrieval, automated discovery of services, etc. Therefore, ontologies and agent-based systems are two different but complementary technologies where the ontology is responsible for providing knowledge to the system while the agents provide dynamism that the system needs [Hadzic et al. 2009b].

Our application corresponds to a family group with an elderly man living alone called *Joao*. He has health problems and needs constant monitoring to perform their daily tasks. *Joao* has two children called *Paulo* and *Stefano*. *Paulo* lives next door with his wife *Jane* and their two children (*Pedro* 12 years old and *Maria* 14 years old). *Stefano* lives in the same city, but about 10 kms away from *Joao*'s house. To help with daily tasks, *Joao* has two professional carers that help him (one for the day and another at night). *Joao* has a routine of activities that includes walk in the park, physiotherapy, stimulation activities (memory games, for instance), as well as feeding and medicines at specific time. The group's tasks are related with the care of the elderly. Whereas the elderly needs to follow up full-time, then, the group established a routine tasks that starts when the elderly wakes up and extend across the rest of the day. Thus the ontology is designed to represent all aspects of the tasks of daily living (ADL) of elderly and its relationship with the other members of the group.

Each application will require a specific instantiation, we instantiate a health care group. In Figure 4 it is possible to see one instance of task concept and their inferences. This instance is the same demonstrated in the plan library (Figure 3). In this example, you can see the specialization of the *Task* concept, as the relationship that it has with the concepts *Person* (through the property *can-be-executed-by*) and *Location* (*occurs-in*). One of inferences refers to the type *CompositeTask* that occurs because the concept

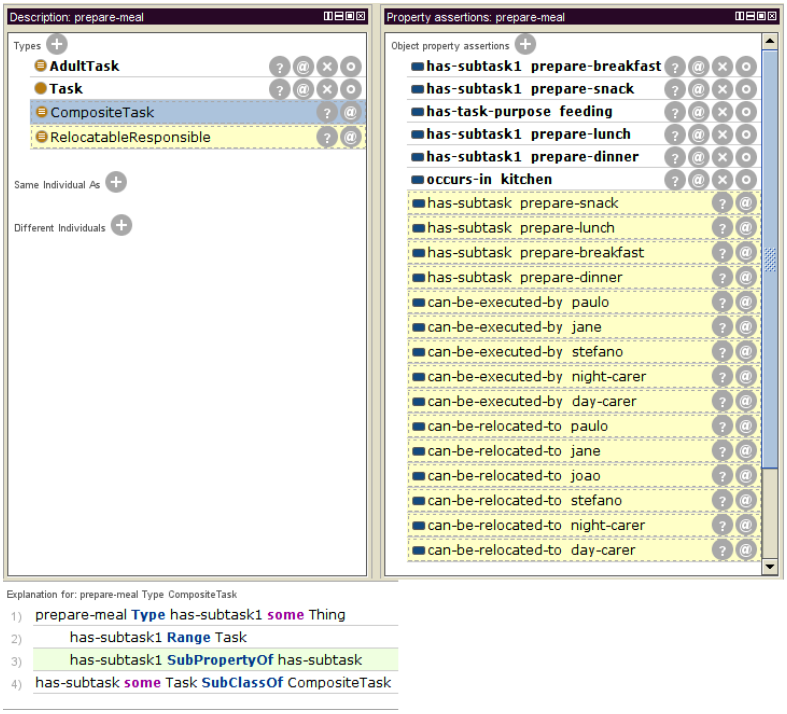

Figure 4. Instance of task concept (*prepare-meal*)

CompositeTask is equivalent to a task that has sub-tasks (the explanation about why the reasoner infers the *CompositeTask* concept can be visualized in Figure 4).

5. Task Ontology Evaluation

We evaluated the task ontology with the group of people that used the ontology as a resource for the Multi-Agent application development. Our goal was to identify if the ontology was considered suitable for their needs in the development. The subjects were: two phd candidates, two masters candidates and one developer. The evaluation consisted of a set of open and closed questions. The closed questions are based on the Likert scale of five points, regarding the following points: *Q1: Do you consider the concepts represented in the ontology relevant for the collaborative multi-agent application?*; *Q2: Do you consider the terminology adequate?*; *Q3: Do you consider that the ontology provides subsidies for the negotiation process?; Q4: Do you consider that the ontology representation is adequate for the plan library generation?*. Table 3 presents the answers. All participants considered it relevant and the terminology was evaluated as adequated. Similarly, the ontology was considered efficient to provide information for negotiation and to the plan library modelling.

In the open questions, the participants could suggest changes in the task concepts (*Q5*), none was suggested. The next questions (*Q6* and *Q7*), ask about advantages and disadvantages of using ontologies in multi-agent systems. Regarding the advantages they mentioned: *(i)* possibility of knowledge reuse; *(ii)* applications development become independent of the domain which may vary according to the ontology and; *(iii)* inferences of new knowledge. Regarding the disadvantages they mentioned: *(i)* performance of the application and *(ii)* the developer needs to know about ontologies.

Table 3. Evaluation of proposed ontology

	Question 1	Question 2	Question 3	Question 4
Strongly Agree	4	4	3	3
Agree	1	0	1	2
Undecided	0	1	1	0
Disagree	0	0	0	0
Strongly Disagree	0	0	0	0

According to the answers, we consider our ontology adequate for the representation of collaborative tasks in multi-agent systems. Issues regarding performance still must be investigated.

6. Final Remarks

This paper presented a new task ontology, based on an extensive literature review of how ontologies are being used as semantic models for task representation and reasoning. We explained how the concepts, properties, and rules were defined in our ontology, and then we exemplified the kind of inference processes that it allows. We have shown how the model allows for knowledge inference about tasks that may be used in the coordination of activities in groups of agents.

The full version of our ontology consists of 34 concepts, 31 object properties, 4 date properties and 73 instances. It allows for different inferences as *(i)* classification tasks into simple and composite, *(ii)* classification tasks that can be performed only by adults, *(iii)* relocation responsible, *(iv)* time constraints, *(v)* classify people as adults or not, among others.

This work is part of a greater research project involving various AI techniques such as MAS development, plan recognition, negotiation, among others. Our focus is the representation of collaborative tasks to provide the required knowledge to other modules developed in the project (more details about the project can be found in our other papers [Freitas et al. 2015], [Panisson et al. 2015]). The ontology was modeled with a level of abstraction that allows it to be reused by other applications whose focus is the representation of collaborative tasks.

The ontology was instantiated allowing reasoning and querying. After we evaluated the ontology by means of a questionnaire that was answered by developers that used the ontology as a source of knowledge for their modules. As future work, we intend to expand the instantiation of tasks to new scenarios of family care and reuse the ontology in the development of other group applications.

Acknowledgment

Part of the results presented in this paper were obtained through research on a project titled "Semantic and Multi-Agent Technologies for Group Interaction", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

Avrahami-Zilberbrand, D. and Kaminka, G. A. (2005). Fast and complete symbolic plan recognition. In *IJCAI-05*, pages 653–658.

Bae, I.-H. (2014). An ontology-based approach to adl recognition in smart homes. *Future Gener. Comput. Syst.*, 33:32–41.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, W3C, http://www.w3.org/TR/owl-ref/.

Chen, L., Nugent, C. D., and 0001, H. W. (2012). A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans. Knowl. Data Eng.*, 24(6):961–974.

Freitas, A., Panisson, A. R., Hilgert, L., Meneguzzi, F., Vieira, R., and Bordini, R. H. (2015). Integrating ontologies with multi-agent systems through CArtAgO artifacts. In *2015 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT*.

Garcia, K., Kirsch-Pinheiro, M., Mendoza, S., and Decouchant, D. (2013). An Ontological Model for Resource Sharing in Pervasive Environments. In *Procedings of International Conference on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*, IEEE/WIC/ACM, pages 179–184.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, pages 907–928. Kluwer Academic Publishers.

Hadzic, M., Wongthongtham, P., Dillon, T., and Chang, E. (2009a). *Ontology-based Multi-Agent Systems*. Springer, Berlin Heidelberg, Germany, 1st edition.

Hadzic, M., Wongthongtham, P., Dillon, T., and Chang, E. (2009b). Significance of ontologies, agents and their integration. In *Ontology-Based Multi-Agent Systems*, volume 219 of *Studies in Computational Intelligence*, pages 93–110. Springer Berlin Heidelberg.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., and Dean, M. (2004). SWRL: A Semantic Web Rule Language combining OWL and RuleML. W3c member submission, World Wide Web Consortium.

Panisson, A. R., Freitas, A., Schmidt, D., Hilgert, L., Meneguzzi, F., Vieira, R., and Bordini, R. H. (2015). Arguing about task reallocation using ontological information in multi-agent systems. In *12th International Workshop on Argumentation in Multiagent Systems (ArgMAS)*.

Riboni, D. and Bettini, C. (2011a). Cosar: Hybrid reasoning for context-aware activity recognition. *Personal Ubiquitous Comput.*, 15(3):271–289.

Riboni, D. and Bettini, C. (2011b). OWL 2 modeling and reasoning with complex human activities. *Pervasive Mob. Comput.*, 7(3):379–395.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: a practical OWL-DL reasoner. *Web Semant.*, 5(2):51–53.

Smith, B. L., Tamma, V. A. M., and Wooldridge, M. (2011). An ontology for coordination. *Applied Artificial Intelligence*, 25(3):235–265.

Tran, V. X. and Tsuji, H. (2007). OWL-T: A Task Ontology Language for Automatic Service Composition. In *ICWS*, pages 1164–1167. IEEE Computer Society.

An Ontology for TNM Clinical Stage Inference

Felipe Massicano[2], Ariane Sasso[1], Henrique Amaral-Silva[1], Michel Oleynik[3], Calebe Nobrega[1], Diogo F. C. Patrão[1]

[1]CIPE - A. C. Camargo Cancer Center

[2]IPEN - USP

{djogo,ariane.sasso,henrique.silva,cnobrega,michel}@cipe.accamargo.org.br

massicano@gmail.com

Abstract. *TNM is a classification system for assessment of progression stage of malignant tumors. The physician, upon patient examination, classifies a tumor using three variables: T, N and M. Definitions of values for T, N and M depend on the tumor topography (or body part), specified as ICD-O codes. These values are then used to infer the Clinical Stage (CS) and reflect the disease progression, which can be 0 (no malignant tumor), IS (in situ), I, II, III, or IV. The rules for inference are different for each topography and may depend on other factors such as age. With the objective of evaluating missing CS information on A. C. Camargo Cancer Center databases, we developed an open ontology to represent TNM concepts and rules for CS inference. It was designed to be easily expansible and fast to compute.*

1. Introduction

Originally developed in 1958 and since then maintained by the Union for International Cancer Control (UICC), the TNM staging system is a cancer classification scheme used mainly to predict survival rates given the disease severity. Based on the fact that patients with localized tumors present higher survival rates when compared to patients with distant metastasis, the TNM staging system aims to help doctors with treatment planning, disease prognosis, interpretation of treatment results and also to facilitate information sharing and improve cancer research [Sobin and Wittekind C 2002].

The classification is based on three main discrete variables: T (0-4), for the evaluation of the primary tumor extension; N (0-3), for the appraisal of the presence and the extension of metastasis in regional lymph nodes; and M (0-1), to annotate the absence or presence of distant metastasis. Some topographies include an additional character in the range $a - d$ for specifying subcategories. Additional characters can also be included to define the information source (clinical exam or pathology biopsy); the diagnosis stage (before/after treatment, after recurrence or through autopsy); and the existence of multiples tumors in the same site. Moreover, other symbols describe optional lymphatic and venous invasion, the histological grade, the metastasis site, presence of isolated tumor cells, sentinel lymph node invasion status, the degree of certainty and the presence of residual tumor after the treatment [Sobin and Wittekind C 2002].

Additionally, each topography has rules for mapping the TNM staging into one variable called clinical stage. The clinical stage ranges from 0 to IV, with an additional character for some sites. Although rules differ for each topography, higher clinical stages

correlates with worse prognosis. Therefore, its determination is a central point in the cancer diagnostic process.

The rules for clinical staging inference, standardized by the TNM staging system, should be used by the physicians during the medical appointment; however, many factors contribute to this not being largely adopted, such as: resistance by physicians to extra paperwork, physicians uncertainty concerning the current staging system and lack of regulatory processes to enforce compliance with the standard [Schmoll 2003]. Many efforts have been made lately to reach that, including its recommendation by specialized medical societies and its use as a mandatory prerequisite for quality accreditation on oncology care [Neuss et al. 2005].

Moreover, the TNM staging information is also crucial for cancer research. As the different clinical stages indicates better or worse response to certain treatments and better or worse prognosis, cancer studies usually focus on diseases of a specific tissue, and a specific clinical stage. If the clinical database does not contain this information for a relevant fraction of the patients, the researchers may have to resort to manually assessing the patient records to find out the sample size.

Since the rules for clinical stage coding are explicitly defined in the TNM publication, it is possible to create a computer program to automatically evaluate them. Such a program would validate existing values, or even provide this information when it is missing. However, representing all rules directly on a computer programming language is an extenuating and repetitive task, and may lead to code maintenance issues. In addition, it would be difficult to a oncology expert, untrained in computer programming, to validate the algorithms.

In order to overcome these difficulties, a proposal to model the concepts, descriptions and rules in TNM clinical stages is to use ontologies. In summary, the term ontology means a specification of a conceptualization and it has been applied to create standardized dictionaries in several fields. [Gruber 1993].

Standardized ontologies have been developed in many areas in such a way that domain experts can share and annotate information in their respective fields. In medicine, well-known standardized and structured vocabularies such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [1], RadLex [Langlotz 2006], Unified Medical Language System (UMLS) [Lindberg et al. 1993], Medical Subject Headings (MeSH) [Nelson et al. 2001] and others have been used for clinical and research purposes. Although new general and specialized ontologies are emerging fast, there is no published ontology yet that approaches the TNM clinical stage coding problem. Yet, some ontologies may represent some of the TNM concepts.

The National Cancer Institute Thesaurus (NCIt) is a reference terminology that covers the clinical care, basic and translational research, public data and also the administrative domain regarding the National Cancer Institute (NCI). It was built upon the NCI Metathesaurus from the UMLS and it is based on description logic with relationships between semantically rich concepts [Smith et al. 2005]. It is coded on OWL Lite, a subset of OWL-DL with enough complexity to represent the ontology data [Bechhofer et al. 2004]. It provides some of the TNM concepts for 6th and 7th edition and each topography has its

[1] `http://www.ihtsdo.org/snomed-ct`

own T, N, M and CS classes with annotations in English. When a concept has the same definition in the 6th and 7th edition, it is defined as a single class, or else specific classes for each version are defined. There is no definition of axioms for inference of Clinical Stage based on values of T, N and M.

The SNOMED CT is a vocabulary comprising more than 310.000 concepts hierarchically organized. There are concepts to represent all TNM (including individual definitions for T, N, M and CS for each topography), however, there are no compositional rules connecting the T, N, M and the topography to the CS. Moreover, its license is not open and there is no official or non-official translation to Portuguese.

Dameron et al. propose the creation of an ontology for automatic grading of lung tumours using OWL-DL description logic language, inspired by the controlled vocabulary for cancer, the NCIt and also by the Foundational Model of Anatomy (FMA) for its anatomical decomposition [Dameron et al. 2006]. Marquet et al. also developed an ontology based on the NCIt for automatic classification of glioma tumors using the WHO grading system. Their ontology contained 243 classes (234 of them corresponding to NCIt classes) which correctly classified simulated tests and graded correctly ten clinical reports out of eleven used on the test for clinical data [Marquet and Dameron 2007]. The links mentioned on both manuscripts for downloading the ontologies were not active at the time of this writing.

The TNM ontology [Boeker et al. 2014] is a thorough representation of the TNM concepts for breast cancer using OWL-DL with SRI expressivity. The focus there was representation of the clinical meaning of each concept: T, N and M, with links to the Foundational Model of Medicine [Rosse and Jr. 2003]. They depict how to represent the tumor, the lymph node, distant metastasis, the organ locations specified and the tumor invasion pattern. Complete as it is, there is no rules for inference of clinical stage, nor the concepts related to the latter.

In this work we present an ontology for allowing inference of the TNM clinical stage of tumors, based on given values of T, N, M, the ICD-O topographic code and other information. This ontology should provide annotations with the original descriptions from the reference, and links to the NCIt ontology wherever applicable.

2. Materials and Methods

The first step was to identify the most common topographies on A. C. Camargo Cancer Center patients. Upon interview with an oncologist expert, we created a list of the ten most relevant topographies for research on this institution. We used the TNM 6th edition, because most of the relevant databases in the institution used this version of the coding system.

To achieve the goal of a fast-computing ontology, we kept its expressivity at the bare minimum while preserving the intended meaning of concepts. We used only subclass, intersection, equivalence, disjunction between classes, and object properties. As seen on Figure 1, the ontology is divided in four files (Figure 1): the main ontology, with the general TNM concepts and the imports of all others; the ICD-O topography, with the topographic classes referred by the TNM; a file with the annotations and finally a file with the clinical stage inference axioms.

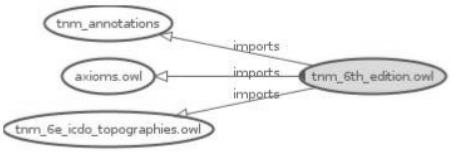

Figure 1. TNM Ontology components and imports diagram.

The concepts for representing T, N, M and CS were created as an hierarchy of classes; the root concept TNM_6th_edition, and its direct subclasses T, N, M and EC (the portuguese acronym for CS). There are subclasses that describes the general classification for all tumors, according to the introduction of the TNM reference. There may be an additional level of subclasses for representing concepts such as T1b or CS IIIa (as defined in some topographies such as breast cancer). We called all those the general staging classes. See Figure 2.

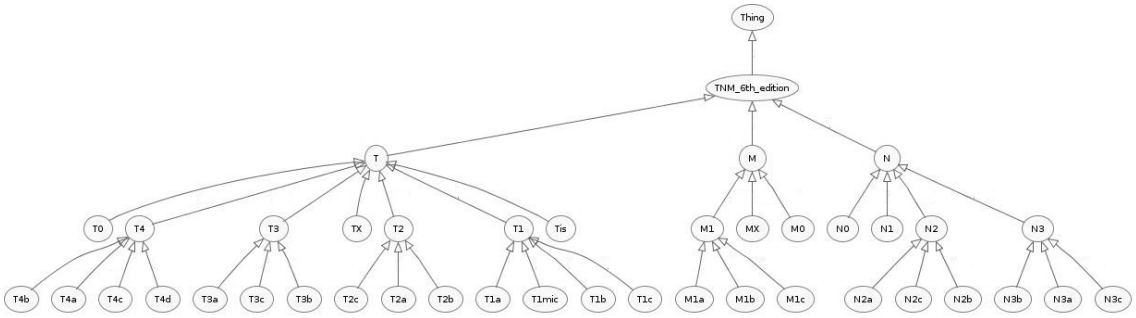

Figure 2. Class hierarchy for TNM concepts.

As the clinical stage rules depends on the tumor topography, the axioms for inference would need reference to ICD-O topography concepts. We could not find any ICD-O ontology available, and it was beyond the scope of our work to create one. However, as ICD-O topographic codes were based on ICD-10 cancer codes, we reused an ICD-10 ontology, available on the BioPortal[2]. We kept only the C00-C80 range of codes, removed some undefined codes within this range (such as C43, C78 and C79) and added C42 (as described in the ICD-O introduction). We also changed the ontology namespace and changed the label annotation property to skos:label. Reference to the prior ontology was kept. In Figure 3 there is a depiction of the ICD-O ontology.

To represent actual patient data, there should be an instance of class Patient, related to one or more instances of class tumor. In order to use the ontology to represent data, an instance representing the tumor should be created and related to subclasses of T, N, M, CS and ICD-O Topography classes. Following the TNM guidelines for staging, a patient with two primary tumors should be represented as one instance of a patient linked to two tumor instances; however, a patient with one tumor that metastasised should have only one tumor instance. The patient instance should be linked to the tumor instances by an object property.

A tumor should not belong to more than one topography class. First, it does not make clinical sense: a tumor should be located on a specific location or organ. It may

[2] `http://bioportal.bioontology.org/ontologies/ICD10`

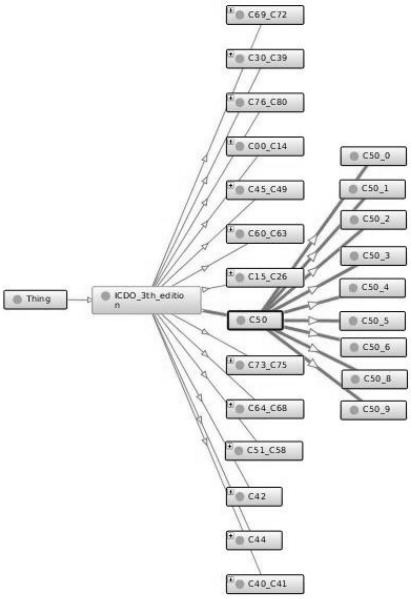

Figure 3. Class hierarchy for ICD-O concepts.

happen to spread itself to neighbour tissues or the precise location maybe be dubious (such as the gastroesophageal junction). In these cases the most probable tumor location should be selected and linked to the instance. The ICD-O Topography ontology states disjunction axioms for all their classes, preventing a tumor instance to belong to two topographic locations at once.

As each topography has different definitions for individual values of the general staging classes T, N and M, we created a script to parse a text file and create a RDF/XML file defining specific staging classes and inference axioms for a pair of T, N or M values and one topography, plus annotations using rdf:Description annotation property. We manually created text files based on the TNM definitions. The axioms are subclasses relating the specific staging classes to the intersection of one general staging class and one topography class.

Whenever a corresponding NCIt concept was available, it was linked to the specific staging class by the property owl:equivalentTo (see Figure 4). Not all concepts defined on TNM were present on NCIt, for instance, the T4 for Breast Cancer.

$$C50 \sqcap M1 \sqsubseteq C50\_M1 \equiv NCIt : C49009$$

Figure 4. Relation between an annotation from the current ontology and a NCIt class.

The standard procedure at the A.C. Camargo Cancer Center is to encode the TNM staging and the ICD-O topography during clinical attendance. As a result, structured information about the clinical stage is not promptly available in its databases. Based on

this, we use the previously constructed inference axioms that considered the values of T, N, M and ICD-O to infer the clinical stage (CS) values.

The format starts with a first line containing the name of the determined clinical stage class. The second line contains one or more topography classes, which are linked to that clinical stage class and separated by a space character. The other lines have a relation of conjunction between the group T, N and M with each specified ICD-O topography. See Figure 5 for an excerpt of these axioms.

$$C50 \sqcap Tis \sqcap N0 \sqcap M0 \sqsubseteq BreastCancer\_CS\_0$$

$$C50 \sqcap T1 \sqcap N0 \sqcap M0 \sqsubseteq BreastCancer\_CS\_I$$

$$C50 \sqcap T2 \sqcap N1 \sqcap M0 \sqsubseteq BreastCancer\_CS\_IIB$$

$$C50 \sqcap N3 \sqcap M0 \sqsubseteq BreastCancer\_CS\_IIIC$$

$$C50 \sqcap M1 \sqsubseteq BreastCancer\_CS\_IV$$

Figure 5. Axioms for inference of clinical stage (CS) based on ICD-O topography and T, N and M classes.

For testing purposes we created another ontology with subjects and patients and assignments to specific classes of this ontology. For each subject we included a topographic class which includes the TNM for each test according to the example below.

$$patientTest00100 : Patient \sqcap hasTumor\ value\ patientTest00100\_Tumor1$$

$$patientTest00100\_Tumor1 : C50 \sqcap Tis \sqcap N0 \sqcap M0$$

After the inference, we can check the TNM annotation classes and also the respective NCIt code class. Thus we reach the ontology objective informing the inferred class to their respective clinical staging. We created a script to generate 566 tests based on the text mappings, as instances of Patient class with exactly one Tumour instance related to it. There were one test for each possible combination of T, N, and other variables for which could be inferred a clinical stage. We created then two queries, one for assessing test instances without any clinical stage inferred (it should have none) and other listing the inferred plus the expected clinical stage for each test.

The software we used to create the ontologies was Protégé [3]. The scripts for the creation of OWL files based on text files were developed in Python. The inferences were computed using Pellet[4].

3. Results

The resulting TNM ontology is divided in four files: main TNM concepts, ICD-O topography, annotations and clinical stage axioms. The main TNM ontology contains the

[3] http://protege.stanford.edu/

[4] https://github.com/complexible/pellet

general staging classes and includes the other ontologies. The ICD-O topography ontology contains the topographic codes and superclasses (such as *C00-C14 - Head and Neck*), with English descriptions. The annotation ontology define the specific TNM classes (such as C50_T1 and C61_M0) and their corresponding description in Portuguese and English. Finally, the clinical stage axioms ontology define the logical axioms that allows the inference of clinical stage based on ICD-O topography and TNM values.

The consolidated ontologies have ALC (Attributive Concept Language with Complements) expressivity. It consisted of 4.382 axioms, 2.954 logical axioms, and 772 classes. It defines 1.690 subClassOf axions, 16 EquivalentTo axioms, 1.248 disjointClasses axioms and 643 AnnotationAssertion axioms. The ontology, the scripts and the text files used to generate it were released under the APACHE-2.0 [5] open source license and are available online at

`https://github.com/djogopatrao/tnm_ontology/tree/master/ontologies`

All 566 test instances were assigned a clinical stage, and only one was assigned two clinical stages. $PatientTest\_51$ was supposed to be assigned Prostate Cancer Clinical Stage I, however an additional concept, Clinical Stage II, was present. This is because the definition of those clinical stages, as stated on the original reference, is ambiguous; Clinical Stage I is defined as T1a, N0, M0 and G1 (Gleason 2-4, discreet anaplasia), while Clinical Stage II, among other definitions, can be T1, N0, M0 and any G. T1, for prostate cancer in the 6th edition of TNM, means "Clinically inapparent tumor neither palpable nor visible by imaging, while T1a (a subconcept for the former) is defined as "Tumor incidental histologic finding in 5% or less of tissue resected. Therefore, as T1a is also T1, so Clinical Stage II is also applicable, and the definition of Clinical Stages in the prostate section of TNM 6th edition contained an ambiguity, detected by means of the ontology.

4. Discussion

We successfully represented the desired TNM rules using an ontology with a simple expressivity profile. That will allow the classification of tumors to remain computable.

The NCIt and SNOMED CT ontologies provide the general concepts involved with tumor staging: the values and description for T, N, M and CS for each topography. However, NCIt does not contains all codes for all topographies. SNOMED CT, in the other hand, does not define which TNM edition their concepts refer to. Neither defined axioms for inferring the clinical stage.

The work by Dameron et al. focus at the anatomical decomposition of a single topography, whereas the present work approaches several topographies, focusing on inference of clinical stage. Besides that, there is no description of the final ontology in the mentioned paper and the links provided are not available [Dameron et al. 2006].

In the paper by Boeker et al, a very detailed description of breast cancer TNM definitions is formalized in a very expressive ontology. The main objective of their work seems to be the formal representation of clinical examination findings for each value of T, N and M, with links to the anatomical and tumoral invasion patterns concepts. That

[5] `http://www.apache.org/licenses/LICENSE-2.0`

allowed the analysis of inconsistencies and inaccuracies in the definitions of TNM itself [Boeker et al. 2014]. However, the ontology at the time of this writing does not include the clinical stage classes, and thus does not provide axioms for their inference. Moreover, this ontology high level of expressivity (SRI) would arguably be less efficient than ALC for a given A-Box.

The tests showed that the inference worked as expected, except in one case, in which the definition provided by the original reference is ambiguous. A related work [Boeker et al. 2014] also found similar ambiguities; this shows how ontologies can be used to prevent classification definition errors.

The presented ontology may be applied to perform validation of existing databases or classify tumors based on TNM values. The usage of relational database to ontology mapping software [Calvanese et al. 2011] [Bizer 2004] [Cullot et al. 2007] allows the usage of the present ontology and inference tools on relational databases, the *de facto* industry standard. As it provides annotations for the meaning of individual T, N and M values for each topography, it may also serve as a reference for physicians and cancer registry workers.

As future work, the presented ontology may be completed to include all topographies and alignment with the NCIt ontology. Alignments with the TNM Ontology [Boeker et al. 2014] may also be of interest. Currently, there are annotations in both Portuguese and English, and other languages may be added. The ontology may be updated to represent the TNM 7th edition, possibly representing an alignment between it and the 6th edition, which may help database migration efforts. Finally, the pathological stage and other modifiers (such as stage post treatment) may also be implemented.

5. Conclusion

We showed that the presented ontology accurately represents the descriptions and inference rules from the selected topographies, fulfilling the main objective of this work. It may be useful in a number of tasks involving tumor staging. It is open source, allowing scrutiny and contributions from the scientific community. It has means to be linked to other TNM ontology efforts and well-established vocabularies, increasing its interoperability. Finally, it is lightweight to compute, being a valuable tool to validate or complete TNM databases.

References

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, W3C, http://www.w3.org/TR/owl-ref/.

Bizer, C. (2004). D2rq - treating non-rdf databases as virtual rdf graphs. In *In Proceedings of the 3rd International Semantic Web Conference (ISWC2004*.

Boeker, M., Faria, R., and Schulz, S. (2014). A Proposal for an Ontology for the Tumor-Node-Metastasis Classification of Malignant Tumors: a Study on Breast Tumors. In Jansen, L., Boeker, M., Herre, H., and Loebe, F., editors, *Ontologies and Data in Life Sciences*, number 1, pages B1–B5, Freiburg.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Ruzzi, M., and Savo, D. F. (2011). The mastro system for ontology-based data access. *Semantic Web Journal*, 2(1):43–53. Listed among the 5 most cited papers in the first five years of the Semantic Web Journal.

Cullot, N., Ghawi, R., and Yétongnon, K. (2007). Db2owl : A tool for automatic database-to-ontology mapping. In Ceci, M., Malerba, D., and Tanca, L., editors, *SEBD*, pages 491–494.

Dameron, O., Roques, E., Rubin, D., Marquet, G., and Burgun, A. (2006). Grading lung tumors using OWL-DL based reasoning. In *9th Intl. Protégé Conference*, pages 1–4, Stanford, California.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220.

Langlotz, C. P. (2006). Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597. PMID: 17102038.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods Archive*, 32(4):281–291.

Marquet, G. and Dameron, O. (2007). Grading glioma tumors using OWL-DL and NCI thesaurus. *AMIA Annual . . .*, pages 508–512.

Nelson, S., Johnston, W. D., and Humphreys, B. (2001). volume 2 of *Information Science and Knowledge Management*, chapter Relationships in Medical Subject Headings (MeSH), pages 171–184. Springer Netherlands.

Neuss, M. N., Desch, C. E., McNiff, K. K., Eisenberg, P. D., Gesme, D. H., Jacobson, J. O., Jahanzeb, M., Padberg, J. J., Rainey, J. M., Guo, J. J., and Simone, J. V. (2005). A process for measuring the quality of cancer care: The quality oncology practice initiative. *Journal of Clinical Oncology*, 23(25):6233–6239.

Rosse, C. and Jr., J. L. M. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478 – 500. Unified Medical Language System.

Schmoll, H.-J. (2003). F.l. greene, d.l. page, i.d. fleming et al. (eds). ajcc cancer staging manual, 6th edition. *Annals of Oncology*, 14(2):345–346.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biol*, 6(5):R46–R46. gb-2005-6-5-r46[PII].

Sobin, L. and Wittekind C (2002). *Classificação de Tumores Malignos*. Wiley and Sons, New York, 6th edition.

OCIP – An OntoClean Evaluation System Based on a Constraint Prolog Extension Language

Cleyton Mário de Oliveira Rodrigues[1,2], Frederico Luiz Gonçalves de Freitas[1], Ryan Ribeiro de Azevedo[1,3]

[1]Center of Informatics – Federal University of Pernambuco, UFPE
50.732-970, Recife-PE, Brazil

[2]FACETEG – University of Pernambuco, UPE
55.294-902, Garanhuns-PE, Brazil

[3]UAG – Rural Federal University of Pernambuco, UFRPE
55.292-270, Garanhuns-PE, Brazil

{cmor,fred,rra2}@cin.ufpe.br

Abstract. *An ontological model must evolve, since several and different knowledge sources can contribute to the addition of new concepts, relations and properties. Hence, we expect a certain level of quality during the engineering of ontologies, as well as the ontological commitment, in order to produce clear, well-formulated and correct subsumption relations. OntoClean is a methodology that addresses the creation of clean ontologies, i.e. the creation of taxonomic hierarchies to model properly the concepts in the domain. Due the lack of stable implementations in the literature, this paper presents OCIP: an OntoClean implementation in Constraint Handling Rules (CHR), a Constraint Programming Prolog extension.*

1. Introduction

Ontologies, in a higher level of abstraction, establish a common and unambiguous terminology for the domain in question. The idea of ontology is often restricted to what is called "formal ontology" [Guarino 1998]. This means that the content of an ontology is described using mathematical logic, which can provide computer system's ability of logical inference. You can also support autonomous discovery from recorded data, as well as reuse and exchange of knowledge. Recently the use of ontologies has been popularized through various other sub-areas of computer science, such as Software Engineering, Database and Information System.

OntoClean [Guarino and Welty 2000] [Welty and Guarino 2001], on the other hand, is a methodology that addresses the creation of clean ontologies, i.e. the creation of taxonomic hierarchies to model properly the concepts in the domain of discourse. OntoClean comprises a set of meta properties, restrictions and assumptions which together defines a methodology for conceptual analysis of taxonomic subsumption (is-a) in any arbitrary ontology. OntoClean does not care about the semantics of the relationship itself, but with the ontological nature of concepts in the relationship. Due to the lack of stable implementations in the literature, this paper presents an OntoClean

implementation in Constraint Handling Rules (CHR^v), a Constraint Programming Prolog extension.

CHR^v [Frühwirth 2009] is a rule based language which was initially conceived to represent white box constraint solvers, but that has been shown to be able to implement many different reasoning services in a straightforward way. Through this language, we have defined a set of rules to check any restrictions violation imposed by OntoClean, known as OCIP (OntoClean Implementation in Prolog).

This paper is organized as follows. Section 2 explores the OntoClean methodology, highlighting the meta properties and restrictions used in this work. Then, Section 3 illustrates CHR^v language through an example for coloring maps. Syntax and Semantics are briefly discussed. Section 4 discusses the implementation of OntoClean in Prolog. Following, Section 5 explains the evaluation of a legal ontology through OCIP, highlighting some violations. Sections 6 explores some related work. Finally, last section presents a conclusion of what has been achieved so far in this research, as well as outlines up prospects for the continuation of this work.

2. OntoClean

An ontological model must evolve, since several and different knowledge sources can contribute to the addition of new concepts, relations and properties. Hence, we expect a certain level of quality during the engineering of ontologies, as well as the ontological commitment, in order to produce clear, well-formulated and correct subsumption relations. That is, decisions regarding the taxonomic structure must faithfully represent the real domain elements and their associations. In addition to leveraging the understanding with a cleaner ontology, the correct establishment of subsumption between these concepts (relying on the ontological nature of them) favors the reuse and integration of these models. Hence, it avoids rework to adjust/tune ontologies by adding new knowledge, allowing them to be widely shared across several information systems.

Therefore, it is suggested that a methodology for decision evaluation is widely required. Furthermore, this methodology must not be directed to a particular domain. It must be general enough to be used and reused in different fields, without adjustments. Thus, this work explores a domain-independent methodology for assessing decisions about the ontological nature of the elements in subsumption relations, namely the OntoClean [Guarino and Welty 2000], [Guarino and Welty 2004]. This methodology, relying on notions arising from philosophical ontology, was proposed by the Ontology Group at the Italian National Research Council (CNR).

OntoClean is a methodology that allows the construction of clean ontologies. Firstly, establishing a set of general and well formalized meta properties so that the concepts can be properly characterized. Secondly, as a result, these meta properties impose a set of constraints between a super and a subclass. Regarding the ontology definition proposed by [Studer et al 1998]: "*A formal explicit specification of a shared conceptualization*", hopefully, through the methodology is possible to detect possible disagreements amongst different conceptualizations, so that some corrective action can be taken.

2.1. OntoClean Meta Properties

Table 1 summarizes the basic notions extracted from philosophical sphere, in which OntoClean is based, namely: Rigidity, Identity, Unity and Dependence. Herein, we will consider concepts and classes as equivalent. Therefore, they represent a collection of individuals, which have been grouped by having common characteristics. The concepts are related by subsumption association, where concept ρ subsumes concept σ, if $\sigma \rightarrow \rho$ which means that all individuals from σ, are also part of ρ, but the reverse is not necessarily true. For further information, the complete basic notions as well as a brief formal analysis can be found at [Welty and Guarino 2001].

Table 1. OntoClean Meta Properties

Meta Property	Symbol	Label	Definition
Rigidity	+R	Rigid	All the instance will always be instances of this concept in every possible world
	-R	Non-Rigid	There are instances that will stop being instances of the concept
	~R	Anti-Rigid	All instances will no longer be instance of that concept
Identity	+I	Carry Identity	Instances carry an unique identification (IC) criteria from any superclass
	-I	Non Carry Identity	There is no identification criteria (IC)
	+O	Supply Identity	Instances themselves provide an unique identification criteria (IC)
Unity	+U	Unity	Instances are “whole”, and have a single unit criteria (UC)
	-U	Non-Unity	Instances are “whole”, but they do not have a single unit criteria (UC)
	~U	Anti-Unity	Instances are not “wholes”
Dependence	+D	External Dependence	There is dependency on external concept
	-D	Non External Dependence	There is no dependency

2.2. OntoClean Constraints

From the methodology, emerges a set of restrictions on the subsumption relations present in the taxonomy. In total, there were defined five restrictions [Guarino and Welty 2000], which follow:

- Anti-rigid class cannot subsume a rigid subclass;
- A class with identity cannot subsume a non-identity subclass;
- A class with the unity meta property cannot subsume a subclass without unity criterion;
- Anti-Unit class cannot subsume unity class;
- Dependent class cannot subsume non-dependent class.

3. CHR^v

Constraint Handling Rules with Disjunction (CHR^v) [Abdennadher and Schütz 1998] is a general concurrent logic programming language, rule-based, which has been adapted to a wide set of applications as: constraint satisfaction [Wolf 2005], abduction [Gavanelli et al 2008], component-development engineering [Fages et al 2008], and so on. The language also emerges as an attempt to integrate an ontological language of the Semantic Web with some rule-based logic programming [Frühwirth 2007]. In essence, it is designed for creation of constraint solvers. CHR^v is a fully accepted logic programming language, since it subsumes the main types of reasoning systems [Frühwirth 2009]: the production system, the term rewriting system, besides Prolog rules. Additionally, the language is syntactically and semantically well-defined [Abdennadher and Schütz 1998].

Without loss of generality, a CHR^v program is a conjunction of simpagation rules, whose syntax is described as follows:

rule_name@ Hk \ Hr <=> G | B.

rule_name@ is the non-compulsory rule identification. The head is defined by the predicates represented by Hk and Hr, with which an engine tries to match with the constraints in the store. Further, G stands for the set of guard predicates, that is, a condition imposed to be verified to fire any rule. Finally, B is the disjunctive body, corresponding to a set of constraints added within the store, whenever the rule fires. The logical conjunction and disjunction of predicates are syntactically expressed by the symbols ',' and ';', respectively. Logically, the interpretation of the rule is as follows:

$$\forall V_{GH}\ (G \rightarrow ((Hk \wedge Hr) \leftrightarrow (\exists\ V_{B\setminus GH}\ B \wedge Hk))),$$

$$\text{where } V_{GH} = vars(G) \cup vars(Hk) \cup vars(Hr),$$

$$V_{B\setminus GH} = vars(B) \setminus V_{GH}$$

For the sake of space, we ask the reader to check the bibliography for further reference to the declarative semantics. Besides the simpagation rule, there are two other cases that are specializations of the former: the simplification (Hr <=> G | B.) which replaces constraints by others equivalent; and the propagation rules (Hk ==> G | B.) which add new constraints within the store, leading to further simplification.

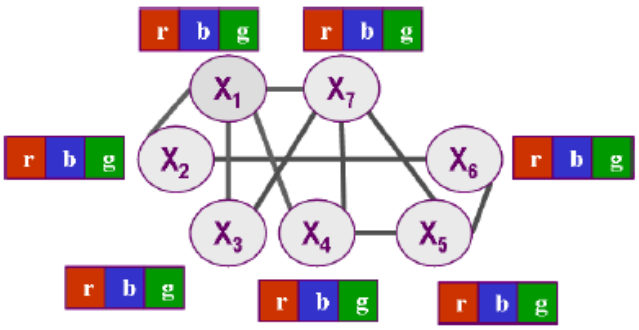

Figure 1: A Pedagogical Map Coloring Problem

Figure 1 illustrates a pedagogical map coloring problem. There are 7 places (X1, X2, X3, X4, X5, X6, X7) whose neighborhood is expressed by an arc connecting these

locations. Further, each one can assume one of the following domain values: D = {r,g,b}, referring to the colors, red, green, and blue, respectively. The only constraint imposed restricts the neighboring places (that is, each pair of nodes linked by an arc) to have different colors. As usual, this problem can be reformulated into a search tree problem, where the branches represent all the possible paths to a consistent solution. By definition, each branch not in accordance with the restriction must be pruned. The problem depicted in Figure 1 is represented by the logical conjunction of the following CHR^v rules:

```
f@ facts ==> m, d(x1,C1), d(x7,C7), d(x4,C4), d(x3,C3), d(x2,C2), d(x5,C5), d(x6,C6).
d1@ d(x1,C) ==> C=red; C=green; C=blue.
d7@ d(x7,C) ==> C=red; C=green; C=blue.
d4@ d(x4,C) ==> C=red; C=green; C=blue.
d3@ d(x3,C) ==> C=red; C=green; C=blue.
d2@ d(x2,C) ==> C=red; C=green; C=blue.
d5@ d(x5,C) ==> C=red; C=green; C=blue.
d6@ d(x6,C) ==> C=red; C=green; C=blue.
m@ m <=> n(x1,x2), n(x1,x3), n(x1,x4), n(x1,x7), n(x2,x6), n(x3,x7), n(x4,x7), n(x4,x5), n(x5,x7),
n(x5,x6).
n1@ n(Ri,Rj), d(Ri,Ci), d(Rj,Cj) <=> Ci=Cj | fail.
```

The first rule `f` introduces the constraints into the store, which is a set of predicates with functor `d` and two arguments: the location and a variable to store the possible color. The seven following rules relate the locations with the respective domain. Additionally, rule `m` adds all the conceptual constraints, in the following sense: `n(Ri,Rj)` means there is an arc linking `Ri` to `Rj`, thus, both places could not share the same color. Finally, the last rule is a sort of integrity constraint. It fires whenever the constraints imposed are violated. Logically, it says that if two linked locations `n(Ri,Rj)` share the same color (condition ensured by the guard), then the engine needs to backtrack to a new (consistent) valuation.

4. OCIP

Due to lack of implementations for OntoClean as already discussed briefly, this paper proposes a new, simple Prolog-based implementation, particularly using the CHR library provided by SWI-Prolog[1]. Being a logic, rule-based and constraint-oriented language, CHR^v has allowed a rapid prototyping of an ontology analyser: OCIP (*OntoClean Implementation in Prolog*). Through propagation rules, a forward chaining reasoner analyzes metaproperties and restrictions pointing out the inconsistencies.

In essence, the analyzer focuses on two logical predicates: `sub/2` and `oc/5`. The former establishes a subsumption relation between two classes of the ontology, i.e. `sub(ClassA, ClassB)` means `ClassA` subsumes `ClassB`. Only direct subsumption relations (between an arbitrary parent class and its immediately children classes) have to be

[1] http://www.swi-prolog.org/

directly defined by the user. The other relations (involving ancestor classes) are trivially propagated through the following transitivity rule:

```
transitivityRule@ sub(CA,CB), sub(CB,CC) ==> sub(CA,CC).
```

Following, the logical predicate oc/5 lists the metaproperties of any class, which is the first argument, and the four other defining respectively the Rigidity, Identity, Unity and Dependence criteria. oc(agent, r, ni, nu, nd) states that the Agent Class is rigid, besides it has non identity, non unity, and non external dependence. Other available labeling criteria are: Anti Rigid (ar), Non Rigid (nr), Identity (i), Owner Identity(o), Unity (u), Anti Unit (au) and finally, Dependence (d).

In order to avoid trivial non-termination (when forward chaining reasoners match indefinitely the same predicates with the same rules), besides the CHR^{v} operational semantics [Duck et al 2004] fully adopted by the SWI-Prolog, some simpagation rules delete equivalent predicates.

```
sympaOcRule@ oc(Class,R,I,U,D) \ oc(Class,R,I,U,D) <=> true.

sympaSubRule@ sub(CA,CB) \ sub(CA,CB) <=> true.
```

In essence, OCIP is based on three blocks rules: rules of natural propagation, horizontal constraints and vertical constraints, plus some auxiliary predicates for explanation of violating constraints. With regard to the rules of natural propagation, new facts contemplating the same class are propagated into the knowledge base. It is known, for example, a class that provide their own identification criterion (+O), is logically a rigid class (+R), which carries an identification criterion (+I). Also, anti-unit classes (~U) are also non-unit classes (-U). The same logical consequence is valid for metaproperties Anti-Ridig (~R) and Non-Rigid (-R).

```
supplyPropagRule@ oc(Class, ,o,X,Y) ==> oc(Class,r,i,X,Y).

unityPropagRule@ oc(Class,X,Y,au,Z) ==> oc(Class,X,Y,nu,Z).

rigidPropagRule@ oc(Class,ar,X,Y,Z) ==> oc(Class, nr,X,Y,Z).
```

Horizontal constraints have this name because they do not analyze superclass/subclass relationships. Unlike, these only evaluate whenever a class has been erroneously characterized with inconsistent metaproperties, like (+R and -R). Therefore, this block has four rules, one for each metaproperty. It is worth noting that anti-properties (~U and ~R) have not been codified, since firing the propagation rules (of the last block), classes should also be classified as -U and -R, respectively.

```
rigidRule@ oc(Class,r, , , ), oc(Class,nr, , , ) ==> rigidViolation(Class).

identityRule@ oc(Class, ,i, , ), oc(Class, ,ni, , ) ==> identityViolation(Class).

unityRule@ oc(Class, , ,u, ), oc(Class, , ,nu, ) ==> unityViolation(Class).

depedenceRule@ oc(Class, , , ,d), oc(Class, , , ,nd) ==> dependentViolation(Class).
```

The 1-ary prolog predicates (rigidViolation, identityViolation, unityViolation, dependentViolation) use other built-in predicates to generate explanations to the user about inconsistencies detected by class. The last rule block corresponds the vertical constraints, that is, those which evaluate the relations of subsumption. For each

OntoClean constraint (mentioned before), a CHR^{v} rule will identify whether there is any violation.

```
antiRigidRule@ oc(ClassSuper,ar, , , ), oc(ClassSub,r, , , ), sub(ClassSuper,ClassSub) ==>
antiRigidViolation(ClassSuper,ClassSub).

noIdentityRule@ oc(ClassSuper, ,i, , ), oc(ClassSub, ,ni, , ), sub(ClassSuper,ClassSub)
==> noIdentityViolation(ClassSuper,ClassSub).

nonUnityRule@ oc(ClassSuper, , ,u, ), oc(ClassSub, , ,nu, ), sub(ClassSuper,ClassSub)
==> noUnityViolation(ClassSuper,ClassSub).

antiUnityRule@ oc(ClassSuper, , ,au, ), oc(ClassSub, , ,u, ), sub(ClassSuper,ClassSub)
==> antiUnityViolation(ClassSuper,ClassSub).

nonDependentRule@ oc(ClassSuper, , , ,d), oc(ClassSub, , , ,nd),
sub(ClassSuper,ClassSub) ==> noDependentViolation(ClassSuper,ClassSub).
```

5. Evaluating OCIP through Legal Ontologies

5.1. Legal Ontologies

OntoCrime and OntoLegalTask [Rodrigues 2015] are ontological representations, through which it is possible to formalize the Brazilian Penal Code[2], making it possible to check the violation of norms, the resolution of legal conflicts (known as antinomies) and the automation of legal reasoning. These formalities have arisen due to semantic deficiencies found in legal texts, either linguistic or conceptual order. OntoCrime emerges as a Domain ontology, defining key concepts and relationships arising from the Penal Code, such as Crime, Punishment, Rules and Articles. OntoLegalTask extends these concepts by defining the tasks mentioned above.

The formalization expected for sound representation and complete reasoning relies on the Descriptions Logic (DLs) [Sirin et al 2007], a decidable subset of First Order Logic (FOL). DLs are the core of OntoCrime and OntoLegalTask representation, for structuring the knowledge bases and for providing reasoning services. Below, we list some DL expressions, which indicate: (i) a person who cannot be criminally punished is the one with a mental illness or a child, teenager or elderly person, (ii) an attributable person is one who does not fit the above profile, (iii) a prohibitive norm prohibits some conduct, (iv) a conduct is prohibited by some article, (v) Crime is a prohibited conduct (vi) with arrest or detention as punishment. For the sake of space, we ask the reader to check the reference for further information.

i. $\text{UnimputablePerson} \equiv \text{NaturalPerson} \sqcap (\exists \text{hasDisorder.MentalDisorder} \sqcup \exists \text{hasAge.(Child} \sqcup \text{Teenager} \sqcup \text{Elderly)}$

ii. $\text{AttributablePerson} \sqsubseteq \neg \text{UnimputablePerson}$

iii. $\text{ProhibitiveArticle} \equiv \exists \text{prohibits.ConductProhibited} \sqcap \forall \text{prohibits.ConductProhibited}$

iv. $\text{ConductProhibited} \sqsubseteq \text{isProhibitedBy.ProhibitiveArticle}$

[2] http://www.planalto.gov.br/ccivil 03/decreto-lei/del2848.htm

v. Crime ⊑ ConductProhibited

vi. Crime ≡ ∃hasPunishment.(Arrest ⊔ Detention)

5.2. Legal Ontologies Labelling

Figures 2 and 3 illustrate a partial view of the the *is-a* relationships extracted from the legal ontologies for analysis. In the legal field conceptualization, Agent is a rigid class whose instances are "whole", but with different unit criteria since the class specializes in Person and Organization. Similarly, instances do not have the same criterion of identity. An instance of an Organization will always be necessarily an organization. Whenever an Organization faces bankruptcy, the entity ceases to be an organization, but also ceases to exist. If an Organization is bought by other, and change the CNPJ[3], the organization also is not the same, it ceased to exist and became a new one. Clearly, the criterion of identity of the instances is the CNPJ itself. A Person, in turn, has as a criterion of identity their own fingerprint.

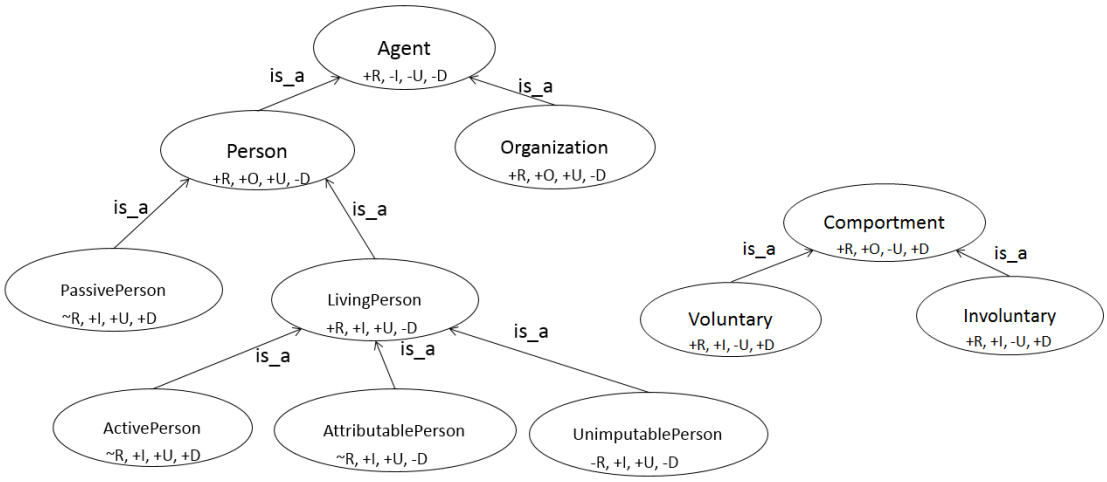

Figure 2: Agent and Comportment subclasses

In criminal law, the passive person is one who suffered the criminal action (or their dependents in the case of murder), while the Active person is the one who practices the act. Thus, a Person may cease to be passive, but it will be the same person. Suppose that this person is an active agent in another crime, notably, it cannot be also passive by the restrictions of the penal code. Another point is that a passive person could be passive in different crimes, but there is not a global criterion of identification; a passive person may be the victim of a thief, or may be the daughter of someone who was murdered, for example. Instances of this class must be related to a criminal conduct, there is an external dependency. The ActivePerson class fits in the same labels.

Class AttributablePerson is anti-rigid: people become old naturally. On the other hand, when a teenager goes into adulthood, he is no longer unimputable. Nevertheless, whether a person has been diagnosed with a mental disorder, even after the legal age, he

[3] CNPJ is the National Register of Legal Entities: a unique number that identifies a legal person

cannot be criminally penalized. Therefore, some instances of UnimputablePerson[4] remain unimputable while live.

With respect to class Comportment (and its subclasses), this depends on the agent who performs the comportment. An instance is bound to be a behavior in all possible worlds. Comportment instances can be identified by a set of variables such as: the action, the agent, place and time. However, there is no single unit criterion: some actions are performed with (criminals) objects, others do not.

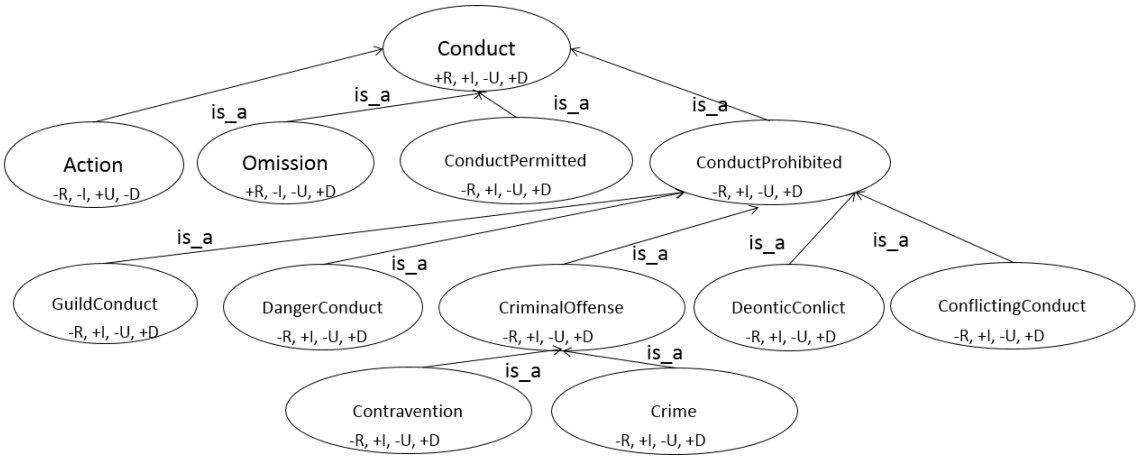

Figure 3: Conduct sublcasses

Similar to Comportment class, Conduct has the same metaproperties, except that it does not provide its own criterion of identity. According to the Criminal Code, Conduct is a voluntary Comportment (thus implicitly Conduct becomes a Voluntary subclass). Among its subclasses, Action class brings together some peculiar characteristics. While a Conduct depends on other external factors, an Action is something more specific, self-contained. Action instances do not share a common IC. Furthermore, their instances have a morphological unit in common (linguistically speaking): are verbs. Finally, the class is labeled with non-rigid metaproperty. Depending on the context, some instances can no longer be part of that concept. When one says, "He walked down the street when he was hit by a car" clearly there is an action. However, when someone says, "He walked nervously", we just have an agent state/condition[5].

ConductProhibited is non-rigid because there may be decriminalization. Instances have an IC in common: the Article prohibitive, with ongoing dependence. Nevertheless, with decriminalization, this prohibitive document will no longer be valid. In principle, the conduct boundaries are known (wholes instances), but there is no single UC for all of them. The ConductProhibited subclasses have the same metaproperties. In GuiltConduct, for example, there are dependencies, as it needs to know the agent

[4] Dead people, animals and legal people cannot be criminally penalized.

[5] In the Portuguese Grammar, in this context, *walk* is a linking verb, which does not indicate action, but a state.

liability. Similarly, for Contravention and Crime, the dependence relates to the kind of penalty imposed by the article of the law. In contrast, due to criminalization, ConductPermitted is labeled as non rigid. The class also has no common IC criteria.

5.3. OCIP Evaluation

In order to carry out the ontological evaluation through OCIP implementation, it was necessary to add into a knowledge base the facts concerning the subsumption relation and the meta properties of the ontology concepts, as previously shown. Some violations were identified as the identity and dependence criteria, between Comportment, Voluntary, Action and Omission classes:

- Identity Class Comportment can not subsume Non Identity Class Action;
- Dependent Class Comportment can not subsume Non Dependent Class Action;
- Identity Class Voluntary can not subsume Non Identity Class Action;
- Identity Class Comportment can not subsume Non Identity Class Omission;

Revisiting the General Theory of Crime within the Brazilian Penal Code, it is said that: "*Conduct **is** any human action or omission, conscious and voluntary, focused on one purpose*". Then, at a first glance, a misinterpretation leads us to believe that a conduct is accomplished through an action or omission, besides being a purely voluntary comportment. Poor written specifications has caused inconsistencies such as these detected by OCIP. Unfortunately, these flaws and ambiguities (besides other linguist and conceptual problems) are present in other legal documents. To fix the inconsistency, it is enough to say that: "*Conduct **has** a human action or omission, conscious and voluntary [...]*". From this perspective, the classes Action and Omission were disconnected as of Conduct specializations. In fact, a Conduct has an action/omission.

6. Related Work

OntOWLClean is a proposal for cleaning ontologies using OWL. The approach has been implemented both in SWOOP tool, as through Protégé [Welty 2006]. In the former case, the tool is no longer available, and in the latter case the plug-in was discontinued. This research had defined two ontologies: one with the metaproperties and restrictions and the other with the semantic definitions to map the classes from a domain ontology in the metaproperties (classes of OntOWLClean). Each new domain ontology then needed to be mapped into OntOWLClean. In addition, the tools have had problems to clearly display or explain the inconsistency.

WebODE[6] was a framework for editing ontologies, which had allowed Ontological analysis, but the environment was discontinued in 2006. Also, the plug-ins available for Ontology edition/evaluation in NeOn project[7] does not support analysis by OntoClean. Being a framework for defining, creating and analyzing, WebODE presented portability issues with other platforms. So an ontology created in WebODE would hardly be analyzed in another tool.

---

[6] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/old-technologies/60-webode

[7] http://www.neon-project.org/

UFO [Guizzardi and Wagner 2004] is a foundational ontology for evaluate business modeling methods, based on other foundational ontologies as OntoClean/DOLCE. OCIP is designed to be a simple tool to assess domain ontologies, even without a foundational ontology. OCIP will be transformed into a plug-in for Protégé that people have a more simple and intuitive interface for analysis. It will be possible for the user, for example, to choose specific metaproperties for analysis, or even graphically shows the backbone of ontology (rigid classes). In fact, the CHR^{v}/Prolog language enables rapid rule prototyping for forward and backward reasoning to manipulate the metaproperties and restrictions. Although, as a future research, we plan to deepen the analysis amongst UFO and OCIP.

6. Conclusion and Future Steps

For a correct and ambiguity-free formalization of knowledge, an important and necessary step is the ontological validation to determine whether there is a close correlation between the domain knowledge and that modeled. OntoClean has emerged as a simple and precise methodology by grouping a set of metaproperties, constraints and assumptions to produce clean and consistent ontological structures. Surprisingly, as far as we know, it has not been identified any valid implementation of OntoClean methodology.

On the other hand, CHR^{v} has become a general and powerful logic language capable of creating constraint-oriented systems through rewriting, propagation, and Prolog-based system. Then this project has followed the idea of creating the OCIP: a Prolog-based implementation (in particular through CHR library), where the OntoClean metaproperties could be attributed, and consequently, the restrictions could be verified. Furthermore, the general purpose CHR^{v} language led to the creation of a simple validator, but that fully covers the methodology proposed by OntoClean.

Our next step will be to make the OCIP implementation plug-in available, so that more and more researchers can use and share their experiences, difficulties and improvements. A key step will be to build a parser able to read the RDF/OWL ontologies code and automatically create the necessary facts (metaproperties and subsumption relationships), leaving the user only labeling directly on the facts of the knowledge base.

Finally, it's worthy of remembering that due to the large amount and the heterogeneity of documents, the ontological evaluation is essential. As soon as more and more models are being built, justify the need for the plug-in to evaluate whether the concepts, their relationships and properties, really express what you see.

References

Abdennadher S. and Schütz, H. (1998) "CHR^{v}, a flexible query language," pp. 1–14.

Baader. F., Calvatese. D., McGuinness. D., Nardi, D. and Patel-Schneider, P. F. (2007) "The description logic handbook, theory, implementation, and applications" (2nd edition). CAMBRIDGE: Cambridge University Press.

Duck, J. D., Stuckey, P. J., de la Banda, M. J. G. and Holzbaur, C. (2004) "The refined operational semantics of constraint handling rules." in ICLP, ser. Lecture Notes in Computer Science, B. Demoen and V. Lifschitz, Eds. Springer, pp. 90–104.

Fages, F., Rodrigues, C. M. O. and Martinez, T. (1998) "Modular CHR with ask and tell," In: CHR '08: Proc. 5th Workshop on Constraint Handling Rules, (Linz, Austria) pp. 95–110.

Frühwirth, T. (2007) "Description logic and rules the CHR way," pp. 49–61, extended Abstract.

Frühwirth, T. (2009) Constraint Handling Rules, 1st ed. New York, NY, USA: Cambridge University Press.

Gavanelli, M., Alberti, M. and Lamma, E. (2008) "Integrating abduction and constraint optimization in constraint handling rules," in Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence. Amsterdam, The Netherlands, The Netherlands: IOS Press, pp. 903–904. [Online].

Guarino, N. (1998) "Formal Ontology in Information Systems". Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy, 1st ed. Amsterdam, The Netherlands, The Netherlands: IOS Press.

Guarino, N. and Welty, C. (2000) "Ontological analysis of taxonomic relationships," in Conceptual Modeling ER 2000, ser. Lecture Notes in Computer Science, A. Laender, S. Liddle, and V. Storey, Eds., vol. 1920.

Guarino, N. and Welty, C. (2004) "An Overview of OntoClean", in Handbook on Ontologies, ser. International Handbooks on Information Systems, S. Staab and R. Studer, Eds.

Guizzardi, G. and Wagner G (2004) A Unified Foundational Ontology and some Applications of it in Business Modelling. Proc on Ws on Enterprise Modelling and Ontologies for interoperability (EMOI-INTEROP).

Rodrigues, C. M. O., Freitas, F. L. G., Silva, E. P., Azevedo, R. R. and Vieira, P. (2015) "An ontological approach for simulating legal action in the brazilian penal code," in Proceedings of The 2015 ACM Symposium on Applied Computing, University of Salamanca, Salamanca, Spain, pp. 376–381.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A. and Katz, Y. (2007) "Pellet: A practical owl-dl reasoner," Web Semant., vol. 5, no. 2, pp. 51–53.

Studer, R., Benjamins, V.R. and Fensel, D. (1998) "Knowledge engineering: Principles and methods," Data Knowl. Eng., vol. 25, no. 1-2, pp. 161–197.

Welty, C. and Guarino, N. (2001) "Supporting ontological analysis of taxonomic relationships," Data Knowledge Engineering, vol. 39, no. 1, pp. 51–74. [Online].

Welty, C. (2006) "OntOWLClean: Cleaning OWL ontologies with OWL," in Proceeding of the 2006 conference on Formal Ontology in Information Systems. Amsterdam, The Netherlands, The Netherlands: IOS Press, pp. 347–359.

Wolf, A. (2005) "Intelligent search strategies based on adaptive constraint handling rules.," Theory and Pratice of Logic Programming 5(4-5), 567-594.

A Systematic Mapping of the Literature on Legal Core Ontologies[1]

Cristine Griffo, João Paulo A. Almeida, Giancarlo Guizzardi

Ontology & Conceptual Modeling Research Group (NEMO),
Computer Science Department, Federal University of Espírito Santo,
Vitória, ES, Brazil

cristine.griffo@aluno.ufes.br, {jpalmeida, gguizzardi}@inf.ufes.br

Abstract*. Over the last decades, the field of legal ontologies has seen a sharp increase in the number of published papers. The literature on legal ontologies now covers a wide variety of topics and research approaches. One of these topics is legal core ontologies, which have received significant attention since the 1990s. In order to provide an up-to-date overview of this research area, this article presents a systematic mapping study of published researches on legal core ontologies. The selected papers were analyzed and categorized according to the perspective of their main contribution as well as according to the legal theories used. The study reveals that only a small number of studies use legal theories suitable to address current societal challenges.*

1. Introduction

The importance of understanding the universe of norms has to do with the broad spectrum of roles that norms play in society. As stated in [Bobbio 2001], individuals, from birth to death, live in a world of norms, which direct their actions. It is thus not surprising that many computer applications are concerned with or manipulate information related to norms, in particular legal norms.

Research in Computer and Law has its roots in the 1960s. In 1957, Mehl [*apud* Bing 2007] wrote about automated legal decisions and initiated a new research trend. Since then, the transdisciplinary area of Computer and Law has matured, with different research niches investigating the various aspects of the field. One of the niches that has received special attention in recent decades is that of *Legal Ontologies*. *Legal Ontologies* is a generic term for ontologies developed to address the legal domain and relates to representation of legal concepts, legal knowledge, and common sense, among others. In contrast, *legal core ontologies (LCO)* are legal ontologies that represent, in the domain of law, domain-independent concepts, properties and relations as well. Applying Guarino's classification of ontologies [Guarino 1998] to the legal domain we can establish the following categories of legal ontologies beyond legal core ontologies: legal domain ontologies, legal task ontologies, and legal application ontologies.

In the early years of research in Computer and Law, researchers did not emphasize the difference between kinds of legal ontologies in their works. The term *legal core ontology* was used in 1996 by [Valente and Breuker 1996] when they

[1] This research is funded by the Brazilian Research Funding Agencies CNPq (grants 311313/2014-0 and 485368/2013-7) and CAPES/CNPq (402991/2012-5). Cristine Griffo is funded by CAPES.

proposed to relate the term *core ontology* used in Van Heijst'thesis (*apud* [Valente and Breuker 1996]) in legal ontology research.

Our investigation of existing *"legal core ontologies"* is motivated by our recent efforts into the construction of a new layer of the Unified Foundational Ontology (UFO) [Guizzardi 2005] in order to represent the legal domain. We started the research based on two pillars to build a consistent legal ontology, as we have defended in [Griffo et al. 2015]: the use of legal theories[2] and foundational ontologies.

Initially, a non-systematic search showed a significant number of papers modeling fundamental legal concepts, such as *claim, duty, obligation and permission*, based on Hohfeld's classification [Hohfeld 1913], [Hohfeld 1917], one of the most important legal theories in the juridical literature. In addition, this preliminary search showed that a noticeable number of papers proposing LCOs had chosen a positivist legal theory as a basis for the ontology, despite the limitations of this particular theory to deal with current legal cases. Finally, it was possible to observe that few legal core ontologies were grounded in a foundational ontology. With this scenario, it was necessary to delimit a study scope and a systematic research method to understand better this field. The scope was limited to *legal core ontologies* and the chosen method was systematic mapping. The genre *legal ontologies* as a whole is not included in the scope, since our focus is to delimit existing work that could in the future inform the design of a unified legal core ontology.

A systematic mapping is an extensive review of primary studies in a specific subject area that aims to identify the available body of work in the literature in order to answer relevant issues [Kitchenham and Charters 2007]. Two points are identified by mapping a specific field of research: 1) difficulties and limitations found by other researchers; and 2) present and future research niches identified by the researchers.

This paper presents the result of systematic mapping of primary studies on *legal core ontologies*, which aimed to:

- Select published studies on legal core ontologies, which mentioned or used either Hohfeld's classification of legal concepts or expressions such as "legal theory" or "legal concepts";
- Classify the selected studies concerning the category of their main contribution: (1) language; (2) tool; (3) method; and (4) model;
- Identify legal theories used in the building of legal core ontologies;
- Identify foundational ontologies used in the building of legal core ontologies;
- Analyze all selected researches in order to point out important research niches into the area of the legal core ontologies.

This paper is outlined as follows: Section 2 presents how the systematic mapping process was developed. In this section, we advance a list of relevant papers as well as the result of their analysis. Section 3 presents final considerations, pointing out the main conclusions of this study, including a discussion on possible bias and limitations.

[2] A legal theory is a body of systematically arranged fundamental principles in order to describe, under a perspective, what exists in the domain of enquiry of the *Law*.

2. The Systematic Mapping Process

In the study reported in this paper, we carried out the systematic mapping process described in [Petersen et al. 2008] and [Kitchenham and Charters 2007], which is illustrated in figure 1. In the first phase of the process, the sources of bibliographical material and both criteria of inclusion and exclusion were defined. Each phase produced an outcome that was used as input for the next phase. As pointed out by [Kitchenham and Charters 2007] and [Petersen et al. 2008], the purpose of a systematic mapping is to provide an overview of a research area in a wide and horizontal way and identify the quantity and type of research and results available within it.

Figure 1. The Systematic Mapping Process. Source: [Petersen et al. 2008] (adapted)

2.1 Planning

In this first phase, both universe and sample of the systematic mapping was delimited. We have chosen *legal ontologies* as the **universe** of our study and *legal core ontologies* as the **sample** of this study. The following questions guided this mapping as well the following inclusion and exclusion criteria.

> **RQ1**: What researches exist in the area of *legal core ontologies?* Which research niches have been investigated (*e.g.* language, method, tool, and model)?
>
> **RQ2:** Which legal theories were used in the selected studies?
>
> **RQ3:** Which foundational ontologies or core ontologies were used on selected legal ontologies?

Inclusion criteria (IC): Papers and chapters of books on legal core ontologies published from 1995 to 2014:

> **IC1**: studies in Computer Science and concerned exclusively with "Computer and Law";
>
> **IC2**: studies that referred to generic legal concepts such as "legal theory" or "legal concepts" (e.g. Hohfeld's classification, legal relation, legal fact).

Exclusion criteria. Exclusion criteria were:

> **EC1**: studies merely available in *abstracts*, slide presentations, technical reports or similar;
>
> **EC2**: duplicity in studies (including versions of the same study, different sources);
>
> **EC3**: studies that were not available in English;
>
> **EC4**: studies about "legal ontologies" only concerned with Law or Philosophy.

Our research source was Google Scholar, which includes papers from different conferences and journals, such as AAAI, ICAIL, JURIX, JURISIN, DEON, RELAW, FOIS, ACM, IEEE, and RuleML, JISCI, Int. Journal of Human-Computer Studies.

2.2 Conduct Search

After the planning phase, the second phase started with the delimitation of the search strings as well as its corresponding control group. Firstly, the *search string* was applied on the sources and the result was compared with the control group in order to minimize a possible bias. The search string was modified to converge with the control group and to include a wider number of studies as well. The outcome of this phase was a total of 269 studies. The selected papers are cited in the text as well as referred to in Section 5.

Search String *= ("legal core ontology" OR "legal core ontologies" OR "legal top-level ontology" OR "generic ontology for legal concepts" OR "core legal ontology" OR "core legal ontologies" OR "legal upper-level ontology" OR "core ontology for law" OR "core ontology of legal concepts" OR "ontology of legal norms" OR core ontology" OR "generic ontology" OR "principled ontology") AND ((Hohfeld OR hohfeldian) OR "legal theory" OR "legal concepts")*

2.3 Screening of the Studies

In this phase, the outcome of phase 2 was refined by considering both inclusion and exclusion criteria. In this phase, we have excluded duplicated studies (found in different sources), technical reports, studies for which a paper with a more recent version had already been included, as well as studies not available in English, or not concerned with ontology in Computer Science. The result of this phase produced a list of 128 selected studies.

2.4 Classification Scheme

Firstly, the outcome of phase 3 was organized by year of publication in order to provide an overview of the LCO area from the chronological point of view (figure 2). Despite that, some papers on legal ontologies have been published since the 1990's; the term *legal core ontology* became more widespread only after the beginning of 2000, peaking in the period of 2005-2009, which sustained attention in the period of 2010-2014.

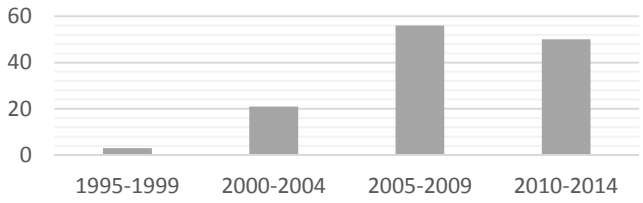

Figure 2. Studies published from 1995 to 2014

A second dimension related to the *contribution perspective* guided the classification of the selected studies in order to identify the major research niches in the area of LCO as well as to identify and analyze three points: 1) the use of legal theories as a theoretical base, 2) the use of foundational ontologies as a base for developing a LCO, and 3) the LCOs encountered in the mapping. For this, we analyzed abstracts, keywords, introduction sections, and references in the selected studies.

2.5 Data Extraction

In this phase, we extracted data from selected papers in order to make a comparative analysis. This analysis consisted of three parts: a) classification and analysis of papers according to their contribution, b) analysis of use of legal theories, and c) analysis of use of foundational ontologies.

Contribution Area. The studies were classified according to their contribution area as shown in figure 3. For this classification, we excluded studies that were either review or opinion papers. The result of this analysis produced a list of 116 studies distributed according to four different contribution areas (see figure 3).

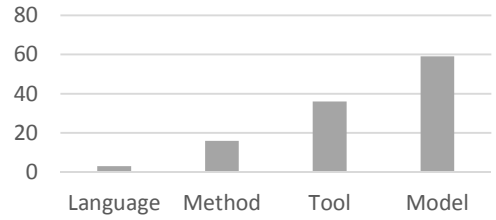

Figure 3. Distribution of researches by contribution perspective

Three among the analyzed papers proposed ***languages*** in legal ontologies. [Athan et al. 2013] presented the LegalRuleML language in the context of the OASIS project and exemplified it with cases of Italian courts. In addition, the Open Digital Rights Language (ODRL), an open standard for expressing machine-readable licenses for digital objects, has been used with ontologies in studies as [García et al. 2005]. Since the scope of this mapping study has temporal and subject boundaries, some studies did not appear. However, it is relevant to point out the following articles, which are related with legal discourse [McCarty 1989], legal relations [Allen and Saxon 1998], legal knowledge [Hamfelt 1995], [Barklund and Hamfelt 1994], and legal argumentation [Gordon 1994].

Regarding the contribution area of ***methods***, the following works have been identified: [Capuano et al. 2014], [Dhouib and Gargouri 2014], [Ceci 2012], [Ceci and Gangemi 2012], [Lenci et al. 2012], [Nguyen and Kaneiwa 2014], [Tiscornia 2010], [Despres and Szulman 2007], [Trojahn et al. 2008], DILIGENT [Casanovas et al. 2005] [Casanovas et al. 2007], TERMINAE [Despres and Szulman 2006] used in [Saravanan et al. 2009] and [Dhouib and Gargouri 2014], Semantic Peer-to-Peer Approach used in EGO ontology [Ortiz-Rodríguez et al. 2006], and Schweighofer's research about legal IR and indexing [Schweighofer 2010].

In the ***tools*** category, we have included applications, systems, databases, and frameworks related with ontologies. Examples in this line of research include: [Hussami 2009], [Drumond and Girardi 2008], [Schweighofer and Liebwald 2007], [Gil et al. 2005], [Moor and Weigand 1999], [Ceci and Ceci 2013], [Lamparter et al. 2005], [Boonchom, V. S., & Soonthornphisaj 2012], [Ceci and Gordon 2012], the FrameNet repository [Venturi et al. 2009], [Venturi et al. 2012], [Breuker et al. 2000], [Wolff et al. 2005], [Kiškis and Petrauskas 2004]. In addition, eGovernance solutions in [Edelmann et al. 2012]; [Tiscornia and Sagri 2012], [Palmirani et al. 2012], [Casellas 2012]; ontology-based application for music digital licenses in [Baumann and Rosnay 2004], [Poblet 2011], [Engers et al. 2008], [Ryan et al. 2003], [Curtoni et al. 1999], [Biasiotti 2011], [Gangemi et al. 2003],[Markovic et al. 2014], DIRECT [Breuker and Hoekstra

2004a], IURISERVICE [Casellas et al. 2007], the LME Project [Bartalesi Lenzi et al. 2009], the LOIS Project [Peters et al. 2006], and a Vietnamese legal application [Thinh et al. 2014].

As shown in figure 3, most of the papers identified in our study, propose particular ***models*** (i.e., *particular ontologies)*. We identified some projects within which ontologies have been developed, such as: DALOS [Agnoloni and Tiscornia 2010], LME [Bartalesi Lenzi et al. 2009], ESTRELLA[3], JUR-IWN or Jur-Wordnet [Casanovas and Poblet and et al. 2005], LOIS [Curtoni et al. 1999], [Tiscornia 2000], [Peters et al. 2006], among others cited. The following legal domain ontologies were build in a project: Medical Law Ontology [Despres and Delforge 2000], Dutch Tax ontology in the E-POWER Project [Boer and Van Engers 2003], International Copyright Law Ontology [Ikeda 2007], Copyright Ontology [García et al. 2007], Mediation Core Ontology (MCO) [Poblet et al. 2009], LAO ontology [Lu et al. 2012], ALLOT ontology [Barabucci et al. 2012], [Despres and Szulman 2004], Ukraine legal ontology [Getman and Karasiuk 2014].

Legal core ontologies. Among legal ontologies found, were found the following legal core ontologies: FOLaw ontology [Valente and Breuker 1994a], [Valente and Breuker 1996], [Valente and Breuker 1994b], Kralingen's ontology [Kralingen 1997], CLO ontology [Gangemi 2007], NM-L+ NM-core ontology [Shaheed, Jaspreet, Alexander Yip 2005], LRI-Core [Breuker and Hoekstra 2004b], Legal-RDF Ontology [McClure 2007], PROTON+OPJK, OPLK [Caralt 2008], Ontological Model of Legal Acts [Gostojic and Milosavljevic 2013], LKIF-core ontology [Hoekstra et al. 2007], [Hoekstra et al. 2009], LegalRuleML-core ontology [Athan et al. 2013] and LOTED core ontology [Distinto et al. 2014].

Use of legal theories. Regarding the *legal theories* referred to selected studies, we focused on identifying the legal theories that were referred to as *primary sources* for the selected works. Among the legal theories, the most cited are: (i) Legal Positivism (appearing in "Pure Theory of Law" [Kelsen 2005], "Some Fundamental Legal Conceptions" [Hohfeld 1913]); (ii) Inclusive Positivism (appearing in "The Concept of Law" [Hart. 1994] and "Norms, Institutions and institutional facts" [MacCormick 1998]);); (iii) Legal Realism (appearing in "Normative System" [Alchourrón, C. E. and Bulygin 1971], "On norms of competence" [Bulygin 1992]); (iv) Legal Interpretivism (appearing in "Taking Rights Seriously" [Dworkin 1978]); (v) Legal Argumentation (appearing in "A Theory of Legal Argumentation" [Alexy 2001], "A Theory of Constitutional Rights [Alexy 2010]", "The New Rhetoric: A Treatise on Argumentation [Perelman, C., Olbrechts-Tyteca 1969]").

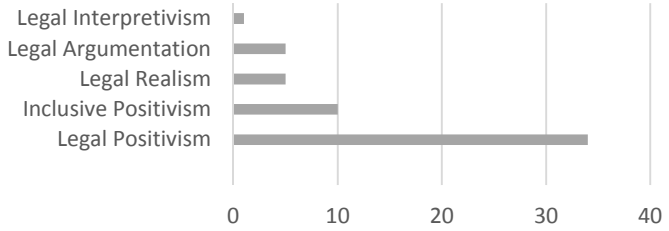

Figure. 4. Main legal doctrines referred in the selected studies

[3] http://www.estrellaproject.org/

The most frequently cited legal doctrines are shown in figure 4. Despite the existence of new legal theories to solve *hard cases*, Legal Positivism is the most frequently used legal theory for legal ontologies. In addition, despite the importance of *legal theory* to legal core ontologies, solely 35 (approx. 27%) of the 128 selected works used primary sources of legal theories; 44 studies (approx. 34%) used indirect sources (*e.g.* use a LCO based on a legal theory to build a domain ontology); and 49 studies (approx. 38%) did not use any primary source.

Use of foundational ontologies. We emphasize the importance of grounding legal ontologies in foundational ontologies in order to obtain ontological quality, as strongly defended by researchers, such as [Guizzardi 2005] and [Uschold and Gruninger 1996]. We identified 47 studies that propose a kind of ontology. Among these ontologies, 32% do not ground the proposed ontology on a foundational/core ontology. The most applied foundational/core ontologies are LKIF, LKIF-core, LRI-CORE, SUMO, DOLCE, CLO, FOLAW and OPJK. Most of the ontologies were specified using OWL, regardless of the various expressiveness limitations (such as those discussed in [Mossakowski et al. 2012]).

3. Final considerations

This paper presents the results of a systematic mapping study investigating published works on the topic of *legal core ontologies*. The systematic mapping revealed that the niche with more extensive literature was the niche of *models*. Some studies used the generic term *"legal ontology"*, giving the idea of a generic or core ontology. An analysis in these studies showed that most of the proposed ontologies were actually *domain ontologies* addressing specific fragments of the Law. In fact, there are few existing LCO, suggesting a research niche to be explored as future work.

This mapping had the purpose of finding existing proposals of LCO in the literature. This purpose was reflected in the *search strings* that we have employed. For this reason, naturally, most of the studies analyzed were cases of studies in which models (ontologies) were proposed. For a more comprehensive research about *language*, *tools* or *methods* in LCO, a change in the search string would be required (with the inclusion of these keywords). For example, visual languages for Law (e.g. Nomos [Ingolfo et al. 2013], [Ingolfo et al. 2014]) suggests an interesting future research topic. Other lines of research that received less attention are tools and, specifically, applications. Finally, the line of research related to *methodologies* in LCO could be explored not only with new proposals of methodologies, but also with evaluation research, qualifying the existing methodologies.

Regarding the issue of legal theories, we select two closely related issues to discuss here. The first point concerns Hohfeld's classification of legal concepts and its meaningful use in the studies of the sample. Exactly 16 studies refer directly to Hohfeld's classification [Hohfeld 1913], [Hohfeld 1917] and many other studies refer to it indirectly. In fact, Hohfeld's classification of legal concepts is one the most important work about classification in Law, and many others theories or classifications are built with Hohfeld's work as a basis. Despite its unquestionable influence in the study of Law *per se*, the popularity of Hohfeld's classification in LCO can also be attributed to its logic-based nature. In the early 20th century, the use of Logic in Law reflected a desire to bring a touch of authority to the discipline. This search for objectivity/scientificity

also contributed to the broad acceptance of Legal Positivism and its doctrine lines at the time. In this context, considering a logic-based approach to the Law as a basis for an LCO, in particular, and computational approaches to the Law, in general, seems like a natural choice. The problem with using theories based on Legal Positivism (*e.g. Hohfeld, Hart, Kelsen theories)* is that they do not include modern concepts of the Law introduced by the explicit countenance of a *social reality*. This problem is propagated to all LCO and computational approaches built following these theories.

The second point is about the (not so) modern theories of Legal Argumentation and Principles (*e.g. Alexy, Perelman, Ryle-Toulmin, Fisher&Patton theories).* In conducting this study, we have observed in recent years an interesting change in Legal-theoretical scenario. The traditional scenario is one in which the so-called "*Purity of the Law*" is sought after, i.e., a scenario in which the analysis of the Law is considered in isolation from the influences of other disciplines such as Economics, Sociology, Anthropology and Politics. We have observed a tendency towards a scenario in which the importance of these related disciplines is acknowledged and openly discussed.

We would like to acknowledge explicitly the following bias in our study. In light of the fact that our study was designed to investigate existing proposals in *"legal core ontologies"*, in our search strings, we did not use the terms "legal ontologies" or "legal ontology" (which were too broad for the scope of this paper). Nevertheless, we are aware that some studies about LCO did use the term *"legal ontologies"* or *"legal ontology"* rather than *"legal core ontology"* or "*legal core ontologies*". In addition, other studies did not use any expressions such as "legal theory", "legal concepts", but rather synonyms (*e.g.* "concept of law"). Examples include: [Gordon 1994], [Visser and Bench-capon 1998], [Visser and Bench-Capon 1996], [Hage and Verheij 1999], [Trojahn et al. [S.d.]], [Allen and Saxon 1998] and [Wyner and Hoekstra 2012].

4. References

Agnoloni, T. and Tiscornia, D. (2010). "Semantic web standards and ontologies for legislative drafting support". In *LNCS*. Springer Berlin.

Alchourrón, C. E. and Bulygin, E. (1971). Normative Systems. *LEP Series*, v. 5.

Alexy, R. (2001). Teoria da argumentação jurídica. 2. ed. Landy Editora.

Alexy, R. (2010). A Theory of Constitutional Rights. Oxford University Press.

Allen, L. and Saxon, C. S. (1998). "The legal argument game of legal relations". In *E-Law-Murdoch U. Electronic J Law*, v. 5, n. 3, p. 1–25.

Athan, T., Boley, H., Governatori, G., et al. (2013)."OASIS LegalRuleML". In *ICAIL'13*. ACM Press.

Barabucci, G., Iorio, A. Di and Poggi, F. (2012). Bridging legal documents , external entities and heterogeneous KBs : from meta-model to implementation". In *SWJ*. IOS Press Journal.

Barklund, J. and Hamfelt, A. (1994). "Hierarchical representation of legal knowledge with metaprogramming in logic". In *The J. Logic Programming*, 18, 1, p. 55–80.

Bartalesi Lenzi, V., Biagioli, C., Cappelli, A., Sprugnoli, R. and Turchi, F. (2009). "The LME project: legislative metadata based on semantic formal models". In *IJMSO 4.3*.

Baumann, S. and Rosnay, M. D. De (2004). Music Life Cycle Support through Ontologies. In *FLAIRS'04 Proceedings*. AAAI Press.

Biasiotti, M. A. (2011). "Semantic Resources for Managing Legislative Information". In: *Legislative XML for the Semantic Web*, p. 151–172. Springer Netherlands.

Bing, J. (2007). "Computers and Law: Some beginnings" In *Information Technology*, 49, 2, p. 71–82.

Bobbio, N. (2001). Teoria da norma jurídica. Edipro, 1ª edição.

Boer, A. and Van Engers, T. (2003). "A Knowledge Engineering Approach to Comparing Legislation". In *KMGov* 2003, LNAI 2645, pp. 139−150. IFIP.

Boonchom, V. and Soonthornphisaj, N. (2012). "ATOB algorithm: an automatic ontology construction for Thai legal sentences retrieval". In *JISCI, 38(1),* p. 37–51.

Breuker, J. and Hoekstra, R. (2004a). "DIRECT:Ontology-based Discovery of Responsibility and Causality in Legal Case Descriptions". In *Jurix 2004*. IOS Press.

Breuker, J. and Hoekstra, R. (2004b). "Core concepts of law: taking common sense seriously". In *FOIS-2004*. IOS Press.

Breuker, J., Petkov, E. and Winkels, R. (2000). "Drafting and Validating Regulations : The Inevitable Use of Intelligent Tools". In *AIMSA'00, LNAI 1904*. Springer Berlin.

Bulygin, E. (1992). "On Norms of Competence". In *Law Philos*, 11, p. 201–216.

Capuano, N., De Maio, C., Salerno, S. and Toti, D. (2014). "A Methodology based on Commonsense Knowledge and Ontologies for the Automatic Classification of Legal Cases". In *WIMS '14*, p. 1–6. ACM New York.

Caralt, N. C. (2008). "Modelling Legal Knowledge through Ontologies. OPJK: the Ontology of Professional Judicial Knowledge".Universitat Autònoma de Barcelona.

Casanovas, P., Casellas, N., Tempich, C., Vrandečić, D. and Benjamins, R. (2007). "OPJK and DILIGENT: ontology modeling in a distributed environment". In *AIL* p. 171–186. Kluwer Academic Publishers.

Casanovas, P, Poblet, M, Casellas, N. et al (2005). "Supporting newly-appointed judges: a legal knowledge management case study". In *JKM*, p.7–27. Emerald Group.

Casellas, N. (2012). "Linked Legal Data: A SKOS Vocabulary for the Code of Federal Regulations". In *SWJ*. IOS Press Journal.

Casellas, N., Casanovas, P., Vallbé, J., et al. (2007)."Semantic Enhancement for Legal Information Retrieval: IURISERVICE performance". In *ICAIL'07*. ACM New York.

Ceci, M. (2012). "Combining ontologies and rules to model judicial interpretation". In *RuleML Workshop Proceedings*.

Ceci, M. and Ceci, M. (2013). "An OWL ontology framework for judicial knowledge". In *Legal Knowledge and Semantic Web framework Conference.*

Ceci, M. and Gangemi, A. (2012). "An OWL Ontology Library Representing Judicial Interpretations". In *SWJ*. IOS Press Journal.

Ceci, M. and Gordon, T. F. (2012). "Browsing case-law: An application of the carneades argumentation system". In RuleML2012@ECAI Challenge, 874, p. 79–95.

Curtoni, P., Dini, L., Di, V., et al. (1999). "Semantic access to multilingual legal information". In Jurix'05. IOS Press.

Despres, S. and Delforge, B. (2000). "Designing medical law ontology from technical texts and core ontology" . In *ECAW 2000 Proceedings*. Springer Berlin.

Despres, S. and Szulman, S. (2004). "Construction of a Legal Ontology from a European Community Legislative Text". In *Jurix'04*, p.79-88. IOS Press.

Despres, S. and Szulman, S. (2006). "Terminae Method and Integration Process for Legal Ontology Building". In *LNCS*, 4031, p 1014-1023. Springer Berlin.

Despres, S. and Szulman, S. (2007). "Merging of legal micro-ontologies from European directives". In *AIL*, 15, 2, p. 187–200. Kluwer Academic Publishers.

Dhouib, K. and Gargouri, F. (2014). "A textual jurisprudence decision structuring methodology based on extraction patterns Arabic legal ontology". In *JDS*, p. 69–81.

Distinto, I., D'Aquin, M. and Motta, E. (2014). "LOTED2 : an Ontology of European Public Procurement Notices". In *SWJ*. IOS Press Jounal.

Drumond, L. and Girardi, R. (2008). "A multi-agent legal recommender system". In *AIL*, pp 175-207. Springer Netherlands.

Dworkin, R. M. (1978). Taking rights seriously. Harvard University Press.

Edelmann, N., Hochtl, J. and Sachs, M. (2012). "Collaboration for Open Innovation Processes in Public Administrations". In *Empowering Open and Collaborative Governance*, p. 21–38. Springer-Verlag Berlin.

Engers, T. Van, Boer, A., Breuker, J. and Valente, A. (2008). "Ontologies in the Legal Domain". In *Digital Government*, p. 233–261. Springer Berlin Heidelberg.

Gangemi, A. (2007). "Design patterns for legal ontology construction". In *Trends in Legal Knowledge*, p.171-191. Europen Press Academic Publishing.

Gangemi, A., Prisco, A. and Sagri, M. (2003). "Some ontological tools to support legal regulatory compliance with a case study". In *OTM 2003 Workshops*. Springer Berlin.

García, R., Gil, R. and Delgado, J. (2007). "A web ontologies framework for digital rights management". In *AIL*, 15, 2, p. 137–154. Kluwer Academic Publishers.

García, R., Gil, R., Gallego, I. and Delgado, J. (2005). "Formalising ODRL Semantics using Web Ontologies". In *ODRL*Ontos'05, p. 137-146. IOS Press.

Getman, A. P. and Karasiuk, V. (2014). "A crowdsourcing approach to building a legal ontology from text". In *AIL*, p.313-335. Springer Netherlands.

Gil, R., Garcia, R. and Delgado, J. (2005). "An interoperable framework for IPR using web ontologies". In *LOAIT*, 15, p. 135–148. IAAIL Press.

Gordon, T. (1994). "The Pleadings Game". In *AIL*, p.239-292. Kluwer Acad.Publishers.

Gostojic, S. and Milosavljevic, B. (2013). "Ontological Model of Legal Norms for Creating and Using Legal Acts". *The IPSI BgD Journal*, 9, 1, p. 19–25.

Griffo, C., Almeida, J. P. A. and Guizzardi, G. (2015). "Towards a Legal Core Ontology based on Alexy's Theory of Fundamental Rights". In *MWAIL, ICAIL 2015*.

Guarino, N. (1998). "Formal Ontology in Information Systems". In *FOIS'98*. IOS Press.

Guizzardi, G. (2005). "Ontological Foundations for Structural Conceptual Model". Universal Press.

Hage, J. and Verheij, B. (1999). "The law as a dynamic interconnected system of states of affairs : a legal top ontology". In *Int. J. Human-Comp St*, 51, p. 1043–1077.

Hamfelt, A. (1995). "Formalizing multiple interpretation of legal knowledge". *In AIL*, 3, 4, p. 221–265. Kluwer Academic Publishers.

Hart, H. (1994). O Conceito de Direito, Fundaçao Calouste Gulbenkian.

Hoekstra, R., Breuker, J., Di Bello, M. and Boer, A. (2007). "The LKIF Core Ontology of Basic Legal Concepts". In *LOAIT'07 Proceedings*, p.43-64. Stanford University.

Hoekstra, R., Breuker, J., Di Bello, M. and Boer, A. (2009). "LKIF core: Principled ontology development for the legal domain". In *LOSW'09*, p. 21-52. IOS Press.

Hohfeld, W. N. (1913). "Some Fundamental Legal Conceptions". In *The Yale Law Journal*, 23, 1, p. 16–59.

Hohfeld, W. (1917). "Fundamental Legal Conceptions as Applied in Judicial Reasoning". In *Faculty Scholarship Series*, paper 4378.

Hussami, L. (2009). "A decision-support system for is compliance management". In *CAISE-DC'09 Proceedings*.

Ikeda, M. (2007). "A uniform conceptual model for knowledge management of international copyright law". In *J Inf Sci*, 34, 1, p. 93–109.

Ingolfo, S., Siena, A. and Mylopoulos, J. (2014). "Goals and Compliance in Nomos 3". In *CAiSE 2014 Proceedings*.

Ingolfo, S., Siena, A., Susi, A., Perini, A. and Mylopoulos, J. (2013). "Modeling laws with nomos 2". In *RELAW*, p. 69–71. IEEE Publisher.

Kelsen, H. (2005). Pure Theory of Law, The Lawbook Exchange, LTD.

Kiškis, M. and Petrauskas, R. (2004)."ICT adoption in the judiciary: classifying of judicial information". In *Int Review Law, Comp Tech*, 18, 1, p. 37–45.

Kitchenham, B. and Charters, S. (2007). "Guidelines for performing Systematic Literature Reviews in Software Engineering". Technical Report, v. 2.3.

Kralingen, R. Van (1997). "A Conceptual Frame-based Ontology for the Law". In *LEGONT'97*, p. 6-17.

Lamparter, S., Oberle, D. and Eberhart, A. (2005)."Approximating service utility from policies and value function patterns". In *POLICY'05*, p. 159–168. IEEE Publisher.

Lenci, A., Montemagni, S., Venturi, G. and Cutrull, M. R. (2012). "Enriching the ISST – TANL Corpus with Semantic Frames". In *LREC'12*, p. 3719–3726.

Lu, W., Xiong, N. and Park, D.-S. (2012). "An ontological approach to support legal information modeling". In *The Journal of Supercomputing*, 62, 1, p. 53–67.MacCormick, N. (1998). "Norms, institutions, and institutional facts". In *Law and Philosophy*, 17, p. 301–345.

Markovic, M., Gostoji, S. and Konjovi, Z. (2014). "Structural and Semantic Markup of Complaints: Case Study of Serbian Judiciary". In *SISY'14 IEEE*.

McCarty, L. T. (1989). "A language for legal Discourse I. basic features. In *ICAIL'89*, p. 180-189.

McClure, J. (2007). "The legal-RDF ontology. A generic model for legal documents". In *LOAIT 2007 Workshop Proceedings*.

Moor, A. and Weigand, H. (1999). "An ontological framework for user-driven system specification". In *HICSS-32'99*, p. 1–10.

Mossakowski, T., Lange, C. and Kutz, O. (2012). "Three Semantics for the Core of Distributed Ontology Language". In *FOIS 2012*. IOS Press.

Nguyen, P. H. P. and Kaneiwa, K. (2014). "Event Inference With Relation and Meta-Relation Type Hierarchies in Conceptual Structure Theory". In *Applied Artificial Intelligence*, 28, 2, p. 139–177.

Ortiz-Rodríguez, F, Palma, R, Villazón-Terrazas, B. Tamaulipeca, U. (2006). "Semantic based P2P System for local e-Government". In *GI Jahrestagung*, p. 329–336.

Palmirani, M., Ognibene, T. and Cervone, L. (2012). "Legal Rules, Text and Ontologies Over Time". In *RuleML@ECAI'12 Proceedings.*

Perelman, C., Olbrechts-Tyteca, L. (1969). The new rethoric: A treatise on Argumentation, University of Notre Dame Press.

Peters, W., Sagri, M.-T., Tiscornia, D. and Castagnoli, S. (2006). "The LOIS Project". In *LREC'06 Proceedings*.

Petersen, K., Feldt, R., Mujtaba, S. and Mattsson, M. (2008). "Systematic mapping studies in software engineering". In *EASE'08 Proceedings*, p. 68–77.

Poblet, M. (2011). "ODR, Ontologies, and Web 2.0". In *Computer*, 17, 4, p. 618–634.

Poblet, M., Casellas, N., Torralba, S. and Casanovas, P. (2009). "Modeling Expert Knowledge in the Mediation Domain: A Middle-Out Approach to Design ODR Ontologies". In *LOAIT'09 Workshop Proceedings*.

Ryan, H., Spyns, P., Leenheer, P. and Leary, R. (2003). "Ontology-Based Platform For Trusted Regulatory Compliance Services". In *OTM 2003 Workshops*, *LNCS 2889*.

Saravanan, M., Ravindran, B. and Raman, S. (2009). "Improving Legal Information Retrieval Using An Ontological Framework". In *AIL*, 17, 2, p. 101–124.
Schweighofer, E. and Liebwald, D. (2007). "Advanced Lexical Ontologies and Hybrid Knowledge Based Systems: First Steps to a Dynamic Legal Electronic Commentary". In *AIL*, p. 103–115. Kluwer Academic Publishers.
Schweighofer, E. (2010) "Semantic Indexing of Legal Documents". In *LNAI 6036*, p. 157-169. Springer Berlin.
Shaheed, Jaspreet, Alexander Yip, and J. C. (2005). "A Top-Level Language-Biased Legal Ontology". In *IAAIL Workshop Series*, v. *4*. Wolf Legal Publishers.
Thinh, B., Quoc, H. and Son Nguyen Truong (2014). "Towards a Conceptual Search for Vietnamese Legal Text". In *CISIM'14*, p. 175-185. Springer Publisher.
Tiscornia, D. (2006). "The LOIS Project: Lexical Ontologies for Legal Information Sharing". In *Legislative XML Worksho*p. European Press Academic Publishing
Tiscornia, D., Agnoloni, T. (2010). "Extracting Normative Content from Legal Texts". In *MCIS'10 Proceedings*.
Tiscornia, D. and Sagri, M. T. (2012). "Legal Concepts and Multilingual Contexts in Digital Information". In *Beijing Law Review*, *3*, *3*, p. 73–80.
Trojahn, C., Quaresma, P. and Vieira, R. (2009). "Matching Law Ontologies using an Extended Argumentation Framework based on Confidence Degrees". In *LOSW*, 188, p. 133-144. IOS Press.
Trojahn, C., Quaresma, P. and Vieira, R. (2007). "Using an Extended Argumentation Framework based on Confidence Degrees for Legal Core Ontology Mapping". In *ArgMAS'07 Proceedings*.
Uschold, M. and Gruninger, M. (1996). "Ontologies: Principles, methods and applications". In *Knowledge Engineering Review*, 11, p. 93–136.
Valente, A. and Breuker, J. (1994a). "A Functional Ontology of Law". In *Artificial Intelligence and Law*, 7, p. 341–361.
Valente, A. and Breuker, J. (1994b). "Ontologies: the Missing Link Between Legal Theory and AI & Law". In *Jurix'94 Proceedings*, p. 138–149.
Valente, A. and Breuker, J. (1996). "Towards Principled Core Ontologies". In *Proc. 10th Workshop on Knowledge Acquisition for Knowledge-Based Systems*.
Van Heijst, G., Schreiber, T. and Wielinga, B. J. (1997). "Using explicit ontologies in KBS Development". In *Int J Hum-Comput St*, p. 183–292.
Venturi, G., Annotated, S., Boella, G., et al. (2012). "Design And Development Of TEMIS: A Syntactically And Semantically Annotated Corpus Of Italian Legislative Texts". In *SPLeT-2012 Workshop Programme.*
Venturi, G., Lenci, A., Montemagni, S., et al. (2009). "Towards a FrameNet Resource for the Legal Domain". *LOAIT'09 Workshop Proceedings*, p. 1–10.
Visser, P. and Bench-Capon, T. (1996). "The Formal Specification of a Legal Ontology. *In Jurix'96 Proceedings*, p. 15–24.
Visser, P. and Bench-Capon, T. J. M. (1998). "A Comparison of Four Ontologies for the Design of Legal Knowledge Systems". In *AIL*, 6, p.27–57. Kluwer Acad. Publishers.
Wolff, F., Oberle, D. and Lamparter, S. (2005). "Economic Reflections on Managing Web Services Using Semantics". *EMISA-2005-Enterprise*.
Wyner, A. and Hoekstra, R. (2012). "A Legal Case OWL Ontology with an Instantiation of Popov v. Hayashi". In *AIL*, 20, 1, p. 83-107. Springer Netherlands.

Supporting FrameNet Project with Semantic Web technologies

Paulo Hauck[1], Regina Braga[1], Fernanda Campos[1], Tiago Torrent[2], Ely Matos[2], José Maria N. David[1]

[1]Pós Graduação em Ciência da Computação
[2]Projeto FrameNet Brasil – Universidade Federal de Juiz de Fora (UFJF)
Campus Universitário s/n – Juiz de Fora – MG – Brazil

{paulo.hauck, regina.braga, fernanda.campos, tiago.torrent, jose.david, ely.matos}@ufjf.edu.br

Abstract. *FrameNet Project is being developed by ICSI at Berkeley, with the goal of documenting the English language lexicon based on Frame Semantics. For Brazilian Portuguese, the FrameNet-Br Project, hosted at UFJF, follows the same theoretical and methodological perspective. This work presents a service-based infrastructure that combines Semantic Web technologies with FrameNet-like databases, by considering the hypothesis that the application of technologies such as ontologies, linked data, and web services can contribute to build and reuse lexical resources based on Frame Semantics. The contributions are related to enriched semantics, data reliability and natural language processing.*

1. Introduction

FrameNet is a lexicography project under development at the International Computer Science Institute (ICSI) with the goal of documenting the English language lexicon based on the concepts from Frame Semantics in [FILLMORE, 1982]. The FrameNet-Br Project is derived from FrameNet, and focuses on the documentation of linguistic frames in Brazilian Portuguese [SALOMÃO, 2011].

There are several works related to the FrameNet Project. Some of them aim to improve data reusability by using technologies that facilitate the reuse of the information contained in the FrameNet database. Among these technologies, one of the most prominent is related to the Semantic Web. The use of Semantic Web technologies emphasizes characteristics such as reuse and acquisition of new knowledge. In the FrameNet context, the Semantic Web can improve the use of lexical data because (i) the formalism provided by ontologies allows formal detailing and definition of shared concepts and the use of inference machines for data validation and implicit information discovery; (ii) the linked data can promote greater integration of FrameNet data with other information bases, like DBPedia and GeoNames and (iii) Web Services allow the integration of tools, independently of both programming languages and operational systems.

On the other hand, the interface between lexical resources and ontologies, the *OntoLex* Interface [Huang et al, 2010], has been recently explored with the aims of understanding how the associations between lexical and formal semantics can contribute to

the improvement of machine reading, an activity that is key to data mining, automatic translation and text summarization.

This paper presents a service-based infrastructure, named FSI (FrameNet Semantic Infrastructure), which combines Semantic Web technologies and FrameNet structure and data. Therefore, this work is related to the benefits that can be obtained with the application of Semantic Web technologies in the context of FrameNet, both in the documentation process and in frame-based searches.

The main objective is to build an infrastructure based on Semantic Web concepts to support the development of FrameNet-like resources, as well as their use and applications. This infrastructure aims to provide two interactive interfaces, one focused on the interaction with other software tools, through a service layer, and another one to support direct user interaction. It allows the maintenance of data, also taking advantage of the benefits of using ontologies for this task.

The specific goals, derived from the main objective are: (i) to provide greater formalism to the FrameNet data, by using ontologies to describe their structures; (ii) to promote the use of FrameNet data by external tools through Web Services; (iii) to provide tools that help in frame documentation and also in sentence annotation; (iv) to reduce the probability of human errors during sentence annotation, using validations based on inference machines; and (v) to provide the user with a new experience on querying FrameNet data, by using linked data and, hence, enabling the discovery of new information. These goals are fully explored by Hauck [2014]. This article particularly focuses on goals (i), (ii) and (v).

This paper is organized into the following sections, besides this introduction. Section 2 briefly presents the main concepts related to frames and the FrameNet Project. Section 3 discusses related work. Section 4 presents the FSI infrastructure and a case study. Finally, section 5 presents the conclusions.

2. Frame Semantics and FrameNet

Frame Semantics proposes that human knowledge is not composed of isolated pieces of information, but is rather based on a set of related concepts. This knowledge is specified in complex structures, called *frames*. These frames constitute a complex system of related concepts so that in order to understand one of them it is necessary to understand the structure in which the entire frame fits [FILLMORE, 1982].

FrameNet [RUPPENHOFER et al, 2011] is a lexical resource for the English language, based on the theory of Frame Semantics. As a lexical resource, it focuses on lexical units, concepts or scenes evoked by theses units (represented by frames), and relations among these frames. The whole project can be seen as an information base, used successfully in applications such as information extraction, machine translation and valence dictionaries. It is also being expanded to other languages such as German[1], Japanese[2],

[1] http://www.laits.utexas.edu/gframenet/

[2] http://jfn.st.hc.keio.ac.jp/

French[3] and Spanish[4]. A version for Brazilian Portuguese has also been developed, called FrameNet-Br [SALOMÃO et al., 2013].

A frame is a structure composed of Frame Elements (FE), which are the participants and props of the scene described by the frame. If a scene is expressed by a sentence, it is said that a specific word in the sentence is the "target", which evokes the frame. Each part of the sentence that is part of the syntactic locality of the target word expresses a Frame Element. The process of defining which part corresponds to each Frame Element is called "annotation" and is, together with frame creation, the main task involved in FrameNet development. According to Ruppenhofer et al. [2010], there are factors that call for the creation of a new frame, such as differences in perspective, variation in the argument structure, causative-inchoative alternation and ontological distinction of FEs. In order to assist the latter factor, FrameNet adopted the definition of Semantic Types for some FEs. The Semantic Type assigned to a FE aims to indicate the type of filler expected to that FE and, on an annotated sentence, one can expect the filler of a FE to be a instance of the assigned Semantic Type.

3. Related Works

Some previous works were discussed during FSI specification, including the use of ontologies to formalize the structure of frames and their relationships [MOREIRA, 2012] [NUZOLESE et al, 2011] [SCHEFFCZYK et al., 2008]; the construction of a service-oriented infrastructure combined with a formal model for the description of their data [VEGI et al, 2011; 2012]; and the development of a tool to support the documentation of frames and the annotation of sentences [LEENOI, 2011].

Scheffczyk et al. [2008] proposes the construction of ontologies in OWL-DL from the transcription of information expressed by FrameNet frames. The ontologies are used to formally describe the structure of a frame. FSI is also based on the idea of creating an ontology to formalize the structure of the frames and their relations, in order to obtain a higher level of data reliability, and to allow other tools to take advantage of these data, considering their formalism. However, we aim to increase the use of ontologies, not only validating the structure of a frame, but also the relationship with other frames and also between FEs. To tackle this issue, we consider that the semantic definition of frames points out that a frame also depends on its relations with other frames, and not only on its components such as FEs and Lexical Units (LU).

In Nuzolese et al. [2011], data from FrameNet were semi-automatically transformed into linked data, using ontologies. According to the authors, this transformation enables greater data integration with other related databases. Similarly to Nuzolese et al. [2011], FSI also uses linked data. However, FSI uses a vocabulary already available in FSI, provided from the data integration, from annotated sentences with data and from other related databases. The advantage of FSI in this case, besides the expressive power of ontologies to define the formal vocabulary of these data, is also in the use of domain

[3] https://sites.google.com/site/anrasfalda/

[4] http://sfn.uab.es:8080/SFN/

ontologies to allow greater expressiveness, as well as the use of external resources connected through linked data, forming a richer knowledge network.

Moreira [2012] revisits some of the limitations that Ovchinnikova el al. [2010] had already pointed out in FrameNet, such as low lexical coverage, incompleteness of the network of relations, inconsistencies in the sets of inherited properties, lack of axiomatization, as well as the fact that FrameNet poses no explicit distinction between roles and types, an important feature for ontologies. Moreira [2012] then proposes that elements of FrameNet structure be formalized so as to avoid mistakes in using them. FSI extends Moreira's [2012] work, by creating an ontology for those elements and also for the data derived from annotation.

Considering the proposal of Leenoi et al. [2011], ontologies were used to formalize part of the data from Thai FrameNet, and they also built tools to support the documentation of frames and the annotation of sentences. For FSI, we also developed tools to support the documentation of frames and the annotation of sentences. Our major differential is that we use semantic information to assist the user in documentation and annotation, ensuring greater data reliability, since, by using inference techniques, the ontology allows the user to notice data inconsistencies.

Vegi et al. [2012] propose an infrastructure for managing and sharing design patterns using metadata descriptions based on a formal vocabulary, and a communication interface to be used by external tools. As Vegi et al. [2012], in FSI formal vocabularies for data representation were created, but with greater expressiveness, by the use of OWL and SWRL rules. In addition, FSI also uses the SOA protocol, thereby promoting greater availability for integration with other tools.

4. FrameNet Semantic Infrastructure

In this section, we present the FSI architecture. FSI is based on SOA principles, and uses Semantic Web concepts together with FrameNet data in order to contribute to the maintenance of FrameNet and the applicability of these data to other activities related to NLP (Natural Language Processing).

Two ontologies were created for FSI implementation: i) FrameNet metadata ontology, named ONTO-FRAME-BR, which semantically describes the data structure that makes up the frames and the semantic relations between them, and ii) ONTO-ANNOTATION-BR, to cover sentence annotation.

FSI aims to reuse existent domain ontologies, which serve as a source for definition of the Semantic Type of Frame Elements. This provides a semantic expressiveness to the fragments of the scene referenced by each Frame Element. The linked data approach [BERNERS-LEE et al, 2001] is also exploited by FSI, for connecting each fragment of a scene, represented by an FE, to a Web resource, so it is possible to get new information from these resources.

4.1 Ontologies

The Copa 2014 FrameNet Brasil Project (COPA2014) [TORRENT et al., 2014] is a frame-based domain specific trilingual electronic dictionary built to be used by tourists, journalists and the staff involved in the organization of the FIFA World Cup 2014 in Brazil.

COPA2014 uses the whole FrameNet infrastructure. We used COPA2014 as a basis to implement and validate FSI. The domain ontologies used for this validation were the PROTON ontology [TERZIEV et al., 2005], which covers various domains but details the tourism domain in depth, and the SWAN Soccer Ontology [MÖLLER, 2004] that covers the soccer domain.

The Onto-Frame-BR aims to provide a semantic basis for the data and metadata. It makes FrameNet data readable by computer engines through the formalism imposed by the ontology. It also contributes to data reliability, since the ontology ensures the semantic validity of the data. To build this ontology, we carried out a reverse engineering process in the COPA2014 project database. The entities that compose the database model, strictly related to the representation of the Frame, were initially mapped as ontology classes. Each relationship between these entities was mapped as object properties. Next, it was necessary to refine the ontology, according to the FrameNet documentation [RUPPENHOFFER et al., 2010]. The first step was to define existential and universal restrictions of classes, in order to validate individuals based on the minimum requirements for their existence. As an example, Figure 1 shows the restrictions for the ontological class `FrameElement`.

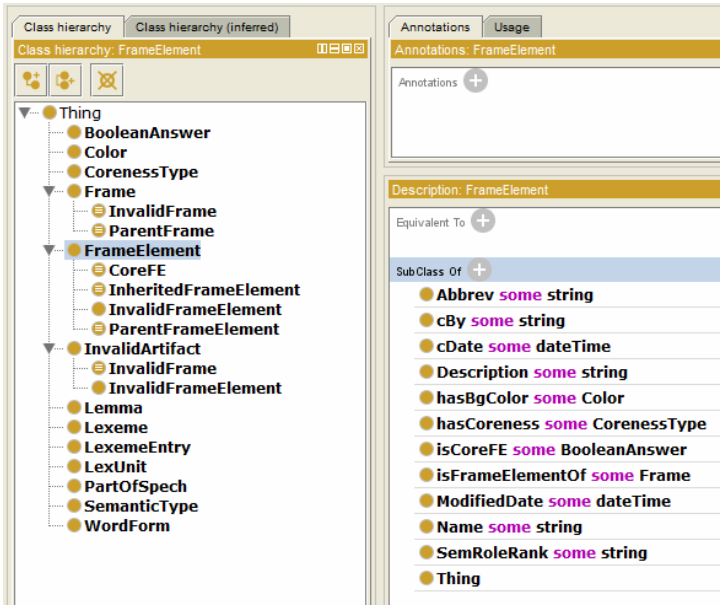

Figure 1: Classes and Restrictions in Protège.

The next step was the separation between frame-to-frame relations and frame internal relations, since in the COPA2014 database, they were grouped together. This separation was made in order to avoid that relations were assigned incorrectly, and also to ensure that the semantic definition of these relations be consistent with that by Ruppenhoffer et al. [2010]. However, some semantic definitions could not be fully specified using only OWL. Thus, SWRL rules were used with the aim of either classifying individuals or identifying implicit relationships that would not be possible only by using OWL.

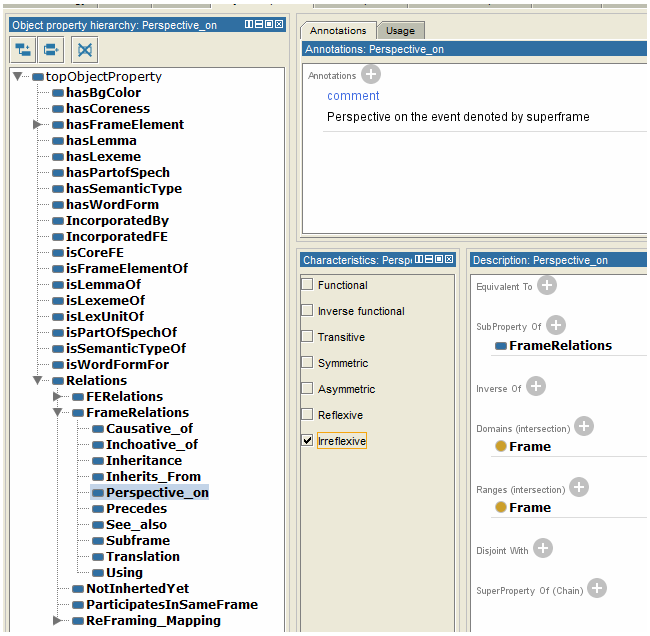

Figure 2: Perspective_on restrictions.

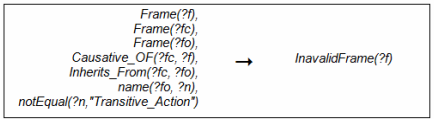

Figure 3: Rule to verify the inheritance of causative frame.

Ruppenhoffer et al. [2010] and Leenoi et al. [2011] describe seven possible frame-to-frame relations and their restrictions. Considering this documentation and in order to adequately represent the structure of FrameNet, the semantics of these relations were defined in FSI. To help in the identification of frames that violate these or any other restrictions defined by SWRL rules, we created a `InvalidFrame` class for those individuals. As an example of one of relations defined in the ontology, we have the `Perspective_on` relation, which is described as a relation between a neutral frame and another non-neutral frame. This relation occurs when a neutral frame can adopt more than one viewpoint. Thus, FEs may vary according to the viewpoint adopted, and the two or more viewpoints can not coexist in the same frame. To explain this restriction, an equivalent property of this relation in the ontology was described as non-reflective, without the need to create SWRL rules (Figure 2).

As an example of SWRL rules creation, we have the `Causative_of` and `Inchoative_of` relations. `Causative` frames should inherit from the `Transitive_action` frame, while `Inchoative` frames should inherit from `Event, State or Gradable_attributes` frames. As shown in Figure 3, the rule for the relation `Causative_Of` checks whether that frame is defined as causative of another frame, and also inherits from a frame that has a different name than `Transitive_action,` so, the rule classifies the target frame from the `Causative_of` relation, in an `Invalid_Frame` class.

Similarly to the frame-to-frame relations, Ruppenhoffer et al. [2010] and Leenoi et al. [2011] also describe possible relations between FEs inside the same frame. In order to support these relations, semantic descriptions in the ontology were also specified. In Figure 4, we can see a summary of all SWRL rules created to support frame internal relations.

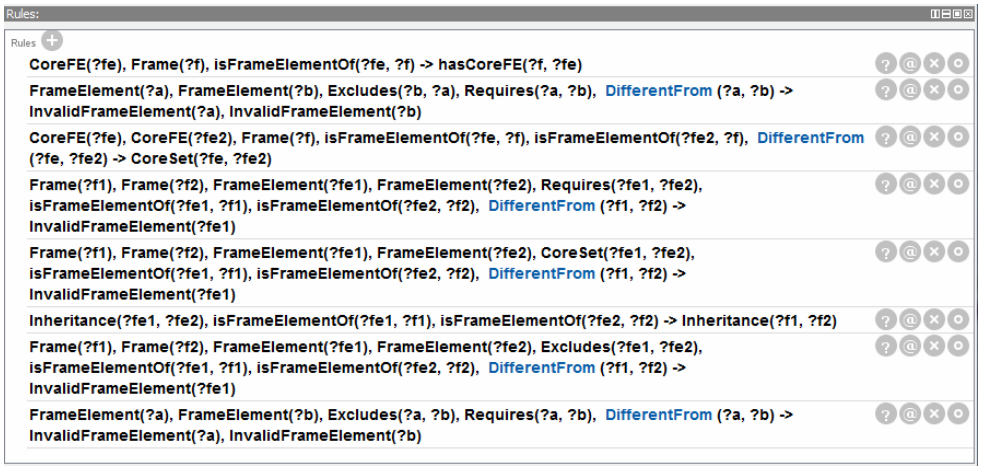

Figure 4: SWRL rules.

As a result of this process, the Onto-Frame-BR was specified. This ontology differs from the ontologies defined in Leenoi et al. [2011], Nuzolese et al. [2011] and Scheffczyk et al. [2008], especially considering the detailed semantics of the relations between frames and between FEs. In Nuzolese et al. [2011] and Scheffczyk et al. [2008] these relations are not expressed or are expressed only as part of the vocabulary without restrictions or rules to validate them. Only in Lenoi et al. [2011] the relations between frames are discussed. But the authors do not make clear if they were treated in the ontology or were only informed. Furthermore, the authors provide no means to obtain or reproduce the ontology.

A partial view of Onto-Frame-BR, presenting its main classes and relations, is shown in Figure 5.

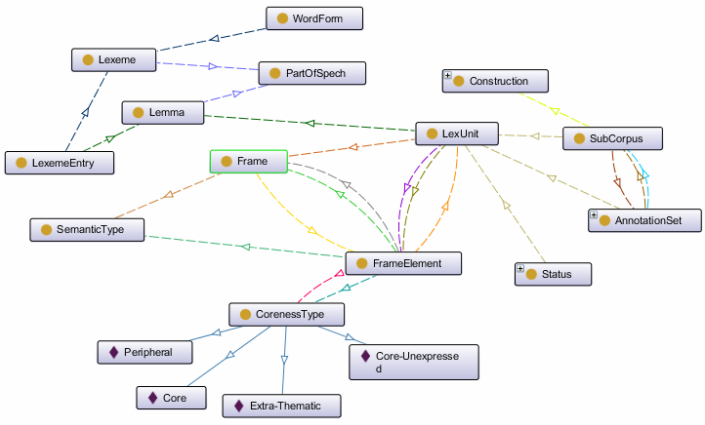

Figure 5: Onto-Frame-BR main classes and relations.

The Onto-Annotation-BR ontology was also developed with the aim of completing the Onto-Frame-BR ontology, covering the semantic annotation, i.e., defining the participation of fragments as FEs and identifying the frame. This ontology allows the representation of annotated sentences carried out in the project. In order to validate the semantics of annotations, two SWRL rules were created, as well as a `InvalidAnnotatedSentence` class for classifying sentences with invalid annotations. Therefore, a way to validate the semantics of annotations was created, using

the two (Onto-Frame-BR and Onto-Annotation-BR) ontologies defined in this work. Furthermore, from the annotated sentences fragments identified in these ontologies, it is possible to associate external linked data resources. These ontologies can be obtained in http://www.ufjf.br/framenetbr-eng/projects/fsi/.

4.2 Architecture

Figure 6 shows an overview of the infrastructure with its main components. FSI is divided into three layers: i) **Data Layer**, where data processed by the infrastructure, such as ontologies, linked data resources, services annotations and access control information, are stored; ii) **Service Layer**, whose purpose is to provide an interface to external software tools (developed in any programming language); and iii) the **Portal**, where an interface is provided. This paper focuses on the description of the **Service Layer**.

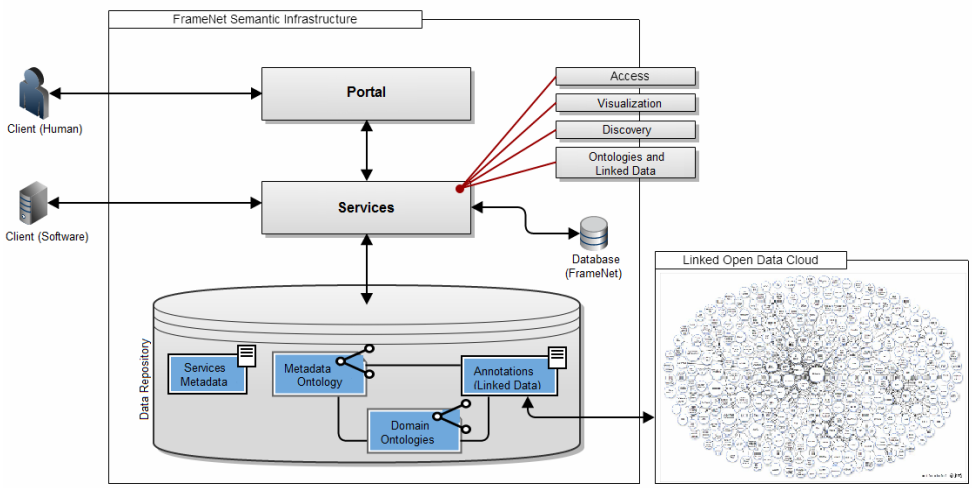

Figure 6: FSI main components.

As stated before, FSI uses a set of ontologies to provide a formal structure and semantics for the data stored in the infrastructure. These ontologies include ONTO-FRAME-BR and ONTO-ANNOTATION-BR, described in section 4.1. The other ontologies are related to the domains that are represented by the frames stored in the database. These domain ontologies allow the definition of semantic restriction on the FEs in a way that makes it possible to evaluate if the annotations respects the semantics of the frame that is evoked. For example, in Figure 7, considering the soccer domain, we have the representation of the frame `Play`, in which their FE `Squads`, `Squad1` and `Squad2` are related to the ontological type `Squad`, which was defined in the Soccer domain ontology. This ontology also defines a restriction where instances of this FE may also be instances of the term `Country` described in the ontology. The same holds for the FE `Host`. However, in this case, `City` and `Country` are both ontological types that can be accepted as an instance of this FE.

For the representation of the fragments that instantiate the FEs, we used linked data sources [Berners-Lee et al, 2001]. Thus, each fragment is connected to at least one term, from an external database, providing more information based on the navigation between these connections. In Figure 8, an example of this approach is presented, considering the annotation "The Brazilian Team faces the USA in Toronto". Where parts of annotations,

such as "The Brazilian Team" and "The USA" are connected by an equivalence relation using a linked data external dataset that represents these teams. Based on these resources, we can get new information from the semantic network that is formed by linked data sets. As an example, we can get the name of the coach or even the names of the players of these teams, taking advantage of the links to external sources.

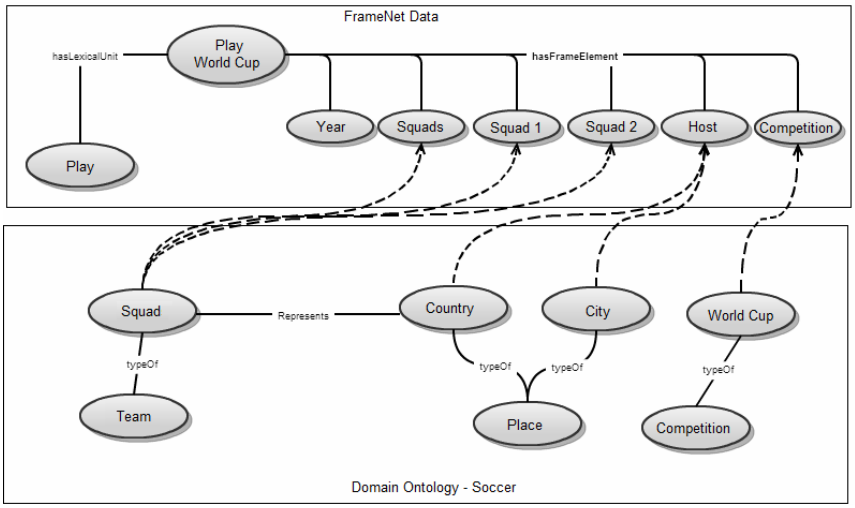

Figure 7: Use of Domain ontologies to restrict the semantic type of FEs.

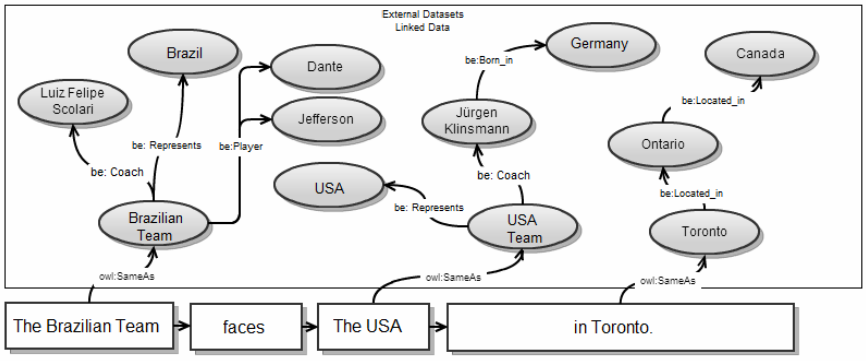

Figure 8: Fragments of annotations using linked data.

The FSI functionalities are available through services, based on SOA architecture. Therefore, four services were developed with the aim of providing a communication interface for external tools: i) **Access Service**: controls the external tools accessing FSI functionalities, avoiding changes in the ontology data; ii) **Visualization Service**: responsible for several data formats that can be provided by the ontology, including the visualization of frames and their structures; iii) **Ontology and Linked Data Service**: responsible for providing an interface to access and modify the ontology data. This service is the most important feature of FSI. It has several methods to obtain FrameNet elements like frames, LUs (lexical units), sentences and annotations. iv) **Discovery Service**: responsible for providing information about services and their methods, including semantic annotations.

4.3. Usage Scenario

In this section we present a usage scenario considering how the interface provided by the Service Layer can be used in NLP activities by external tools.

To illustrate this scenario, we used the Cadmos tool (Character-centered Annotation of Dramatic Media Objects) [CATALDI et al., 2011]. It is a framework to support the annotation of multimedia resources based on the use of ontologies and on the identification of scenes.

During the description of a scene, terms and expressions with ambiguous meanings and different interpretation possibilities may appear. To tackle this issue, Cadmos provides a disambiguation process that uses various lexical resources, including FrameNet and WordNet. In Cadmos, the generated annotations are stored in RDF triples and associated to ontologies for domain delimitation. Since WordNet and FrameNet are supported for scene identification, FSI may be used in the frame disambiguation process, as shown in Figure 9. One of the advantages of using FSI in this context is the use of semantic information that can be obtained from FEs, since these elements may be assigned to the domain ontology, making possible to better identify the context in which the frame can be applied. In addition, it could also be possible to take advantage of FrameNet annotation data, stored in FSI, which are associated with external linked data sources. These connections can enrich the media annotations, for example, by assigning an annotated sentence element from FrameNet to a Cadmos annotation element.

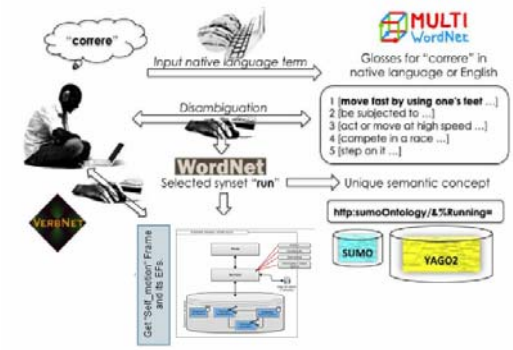

Figure 9: Terms Disambiguation process in Cadmos with FSI frames data.

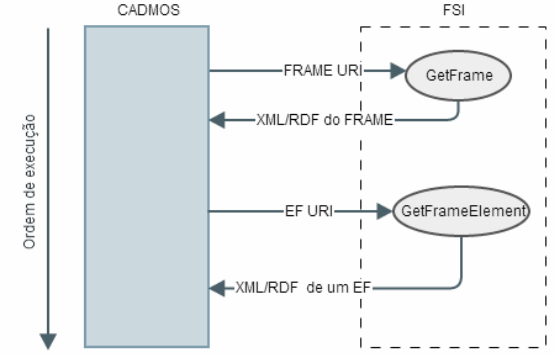

Figure 10: Service Flow Execution in order to obtain FSI frames and FEs data.

Figure 10 details the interaction flow between Cadmos and the methods of the ontology and linked data service of FSI to obtain the frames and their FEs data in the disambiguation process.

5. Conclusions

Several authors have been contributing to improve the access to lexical resources such as FrameNet, as well as their use in different applications and the sharing of related information. Those efforts benefit from Semantic Web technologies, such as ontologies and linked data. These technologies, applied to FrameNet, can provide formalization of frame structure using both formal vocabularies and ontological classes.

This work follows this approach by combining: i) the use of ontologies that describe the structure of frames and semantic relations between these frames associated with the use of domain ontologies for semantic constraints of FEs; ii) the use of linked data to enrich the annotation of sentences; and iii) the access to data through a Service Layer that enables the integration of FSI with other services and applications.

The main contributions of the work are: i) the construction of an infrastructure, based on Semantic Web and SOA technologies, to foster the access to lexical resources and to promote more reliability to the documentation of frames and annotation of sentences; ii) the construction of ONTO-FRAME-BR, which formally represent the frame structure and deals with the semantics of the relations between frames and between their elements, supporting the frame documentation process and providing the user with evidence of possible errors; iii) the construction of ONTO-ANNOTATION-BR, which helps structure the process of sentence annotation so that sentence fragments can be both related to FEs documented in ONTO-FRAME-BR and used as linked data; iv) the possibility of using domain ontologies to relate external linked data resources to fragments of annotated sentences.

Some limitations may also be highlighted, both related to the technology and to the scope adopted. Among them, we list: i) the limitations of OWL and SWRL to treat inheritance relations between frames; ii) the fact that only the semantic aspects of sentence annotation were accounted for in FSI. .

Despite these points to be improved, we believe that the work achieved its objectives by providing an infrastructure that contributes to FrameNet both in regards to maintenance issues and to the offering of semantic information that can be used by external users and tools.

Acknowledgments

We would like to thanks FAPEMIG, CNPq and CAPES for their support.

References

BERNERS-LEE, T., HENDLER, J., LASSILA, O. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 01/05/2001. Disponível em: <http://www.librarything.com/work/2389150>. Acesso em 16 abril 2013.

CATALDI, M., DAMIANO, R., LOMBARDO, V., PIZZO, A., SERGI, D. Integrating commonsense knowledge into the semantic annotation of narrative media objects. In: Artificial Intelligence Around Man and Beyond. Springer Berlin Heidelberg, 2011.

FILLMORE, C. J. Frame Semantics. In: The Linguistic Society Of Korea (org.). Linguistics in the morning calm. Seoul: Hanshin, 1982.

HAUCK, P. FSI: Uma infraestrutura de apoio ao Projeto Framenet utilizando Web Semântica. 142p. Dissertação de Mestrado em Ciência da Computação. Universidade Federal de Juiz de Fora, 2014.

HUANG, C. et al. Ontology and the Lexicon. Cambridge, MA: Cambridge University Press, 2010.

LEENOI, D., JUMPATHONG, S., PORKAEW, P., SUPNITHI, T. Thai Framenet Construction and Tools. International Journal on Asian Language Processing 21(2), p. 71-82. 2011.

MOREIRA, A. Proposta de um framework apoiado em ontologias para a detecção de frames. 2012. 194p. Tese de Doutorado em Linguística. Universidade Federal de Juiz de Fora, 2012;

MÖLLER, K. SWAN Soccer Ontology. 2004. Disponível em http://sw.deri.org/2005/05/swan/soccer/ontology/soccer.owl. Acessado em: 12 Set. 2013.

NUZZOLESE, A. G., GANGEMI, A., PRESUTTI, V. Gathering lexical linked data and knowledge patterns from FrameNet. In: Proceedings of the sixth international conference on Knowledge capture (K-CAP '11). ACM, New York, USA, p 41-48. 2011.

OVCHINNIKOVA, E. *et al.* Data-driven and ontological analysis of FrameNet for natural language processing. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). ELRA. Valletta, Malta.

RUPPENHOFER, J., ELLSWORTH, M., PETRUCK, M. R. L., JOHNSON, C. R., SCHEFCZYK, J. FrameNet II: extended theory and practice. 2010. Disponível em: <http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>. Acesso em: 12 Jan 2013.

SALOMÃO, M. M. M, TORRENT, T.T., SAMPAIO, T. F. A Linguística Cognitiva Encontra a Linguística Computacional: notícias do projeto Framenet Brasil. Cadernos de Estudos Linguísticos 55 (1), p. 7-32. 2013. (in portuguese)

SCHEFFCZYK, J., BAKER, C. F., NARAYANAN, S.. Ontology-Based reasoning about lexical resources. In Ontologies and Lexical Resources for Natural Language Processing, Cambridge Studies in Natural Language Processing. Cambridge University Press, Cambridge. 2008.

TERZIEV, I., KIRYAKOV, A., MANOV, D. Base upper-level ontology (BULO) Guidance. Deliverable of EU-IST Project IST. 2005.

TORRENT, T. T., SALOMÃO, M. M. M., CAMPOS, F. C. A., BRAGA, R. M. M., MATOS, E. E. S., GAMONAL, M. A., GONÇALVES, J. A., SOUZA, B. C. P., GOMES, D. S., PERON, S. R. Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin, Ireland, p 10-14. 2014.

VEGI, L. F. M. Technical description of Dublin Core application profile to analysis patterns (DC2AP). 2012. Disponível em: <http://purl.org/dc2ap/ TechnicalDescription>. Acesso em: 22 Jan. 2013.

VEGI, L. F. M., PEIXOTO, D. A., SOARES, L. S., LISBOA FILHO, J., OLIVEIRA, A. P. An infrastructure oriented for cataloging services and reuse of analysis patterns. In: International Workshop On Reuse In Business Process Management, 2, 2011, Clermont-Ferrand, France. Proceedings of BPM 2011 Workshops, LNBIP vol. 100, Part 4. Berlin: Springer, 2012. p. 338-343.

Conceiving a Multiscale Dataspace for Data Analysis

Matheus Silva Mota[1], **André Santanchè**[1]

[1]Institute of Computing
UNICAMP
Campinas – SP – Brazil

{mota,santanche}@ic.unicamp.br

Abstract. *A consequence of the intensive growth of information shared online is the increase of opportunities to link and integrate distinct sources of knowledge. This linking and integration can be hampered by different levels of heterogeneity in the available sources. Existing approaches focusing on* heavyweight *integration – e.g., schema mapping or ontology alignment – require costly upfront efforts to handle specific formats/schemas. In this scenario, dataspaces emerge as a modern alternative approach to address the integration of heterogeneous sources. The classic heavyweight upfront one-step integration is replaced by an incremental integration, starting from* lightweight *connections, tightening and improving them when benefits worth such effort. Based on several previous work on data integration for data analysis, this work discusses the conception of a multiscale-based dataspace architecture, called LinkedScales. It departs from the notion of integration-scales within a dataspace, and defines a systematic and progressive integration process via graph-based transformations over a graph database. LinkedScales aims to provide a homogeneous view of heterogeneous sources, allowing systems to reach and produce different integration levels on demand, going from raw representations (lower scales) towards ontology-like structures (higher scales).*

1. Introduction and Motivation

From science to business, several domains are facing a huge increase in the amount of available data and the growth of the data heterogeneity (in various levels). In parallel, opportunities may emerge from the exploitation of the increasing volume of connections among multidisciplinary data [Hey et al. 2009].

Domains like biology are increasingly becoming data-driven. Although they adopt different systems to produce, store and search their data, biologists increasingly need a unified view of these data to understand and discover relationships between low-level (e.g., cellular, genomic or molecular level) and high-level (e.g., species characterization, macro-biomas etc.) biological information among several heterogeneous and distributed sources. Therefore, integration becomes a key factor in such data-intensive and in multidisciplinary domains; the production and exploitation of connections among independent data-sources become essential [Elsayed and Brezany 2010]. Besides integration, challenges like provenance, visualization and versioning are experienced by domains that handle large, heterogeneous and cross-connected datasets [Heath and Bizer 2011].

In order to integrate available sources, classical data integration approaches, found in the literature, usually require an up-front effort related to schema recognition/mapping

in an all-or-nothing fashion [Halevy et al. 2006a]. On demand integration of distinct and heterogeneous sources requires ad hoc solutions and repeated effort from specialists [Franklin et al. 2005].

Franklin et. al propose the notion of *dataspaces* to address the problems mentioned above [Franklin et al. 2005]. The dataspace vision aims to provide the benefits of the classical data integration approach, but via a progressive "pay-as-you-go" integration [Halevy et al. 2006a]. They argue that linking lots of "fine-grained" information particles, bearing "little semantics", already bring benefits to applications, and more links can be produced on demand, as *lightweight* steps of integration.

Related work proposals address distinct aspects of dataspaces. Regarding the architectural aspect, each work explores a different issue of a dataspace system. Among all efforts, no dominant proposal of a complete architecture has emerged until now. We observed that, in a progressive integration process, steps are not all alike. They can be distinguished by interdependent roles, which we organize here as abstraction layers. They are materialized in our LinkedScales, a graph-based dataspace architecture. Inspired by a common backbone found in related work, LinkedScales aims to provide an architecture for dataspace systems that supports progressive integration and the management of heterogeneous sources.

LinkedScales takes advantage of the flexibility of graph structures and proposes the notion of scales of integration. Scales are represented as graphs, managed in graph databases. Operations become transformations of such graphs. LinkedScales also systematically defines a set of scales as layers, where each scale focuses in a different level of integration and its respective abstraction. In a progressive integration, each scale congregates homologous lightweight steps. They are interconnected, supporting provenance traceability. Furthermore, LinkedScales supports a complete dataspace lifecycle, including automatic initialization, maintenance and refinement of the links.

This paper discusses the conceiving of the LinkedScales architecture and is organized as follows. Section 2 discusses some concepts and related work. Section 3 introduces the LinkedScales proposal, also discussing previous work and how such experiences led to the proposed architecture. Section 4 presents previous work on data integration and discusses how such experiences are reflected in current proposal. Finally, Section 5 presents some conclusions and future steps.

2. Related Work

2.1. The Classical Data Integration

Motivated by such increasingly need of treating multiple and heterogeneous data sources, data integration has been the focus of attention in the database community in the past two decades [Hedeler et al. 2013]. One predominant strategy is based on providing a virtual unified view under a global schema (GS) [Kolaitis 2005]. Within GS systems, the data stay in their original data sources – maintaining their original schemas – and are dynamically fetched and mapped to a global schema under clients' request [Lenzerini 2002, Hedeler et al. 2013]. In a nutshell, applications send queries to a mediator, which maps them into several sub-queries dispatched to wrappers, according to metadata regarding capabilities of the participating DBMSs. Wrappers map queries to the

underlying DBMSs and the results back to the mediator, guided by the global schema. Queries are optimized and evaluated according to each DBMS within the set, providing the illusion of a single database to applications [Lenzerini 2002].

A main problem found in this "classical" data integration strategy regards the big upfront effort required to produce a global schema definition [Halevy et al. 2006b]. Since in some domains different DBMSs may emerge and schemas are constantly changing, such costly initial step can become impracticable [Hedeler et al. 2013]. Moreover, several approaches focus on a particular data model (e.g., relational), while new models also become popular [Elsayed et al. 2006]. As we will present in next section, an alternative to this classical all-or-nothing costly upfront data integration strategy is a strategy based on progressive small integration steps.

2.2. The "Pay-as-you-go" Dataspace Vision

Since upfront mapping between schemas are labor intensive and scheme-static domains are rare, pay-as-you-go integration strategies have gained momentum. Classical data integration (presented in Section 2.1) approaches work successfully when integrating modest numbers of stable databases in controlled environments, but lack an efficient solution for scenarios in which schemas often change and new data models must be considered [Hedeler et al. 2013]. In a data integration spectrum, the classical data integration is at the high-cost/high-quality end, while an incremental integration based on progressive small steps starts in the opposite side. However, this incremental integration can be continuously refined in order to improve the connections among sources.

In 2005, *Franklin et. al* published a paper proposing the notion of *dataspaces*. The dataspace vision aims at providing the benefits of the classical data integration approach, but in a progressive fashion [Halevy et al. 2006a, Singh and Jain 2011, Hedeler et al. 2010]. The main argument behind the dataspace proposal is that, in the current scenario, instead of a long wait for a global integration schema to have access to the data, users would rather to have early access to the data, among small cycles of integration – i.e., if the user needs the data now, some integration is better than nothing. This second generation approach of data integration can be divided in a bootstrapping stage and subsequent improvements. Progressive integration refinements can be based, for instance, on structural analysis [Dong and Halevy 2007], on user feedback [Belhajjame et al. 2013] or on manual / automatic mappings among sources – if benefits worth such effort.

Dataspaces comprise several challenges related to the design of Dataspace Support Platforms (DSSPs). The main goal of a DSSP is to provide basic support for operations among all data sources within a dataspace, allowing developers to focus on specific challenges of their applications, rather than handling low-level tasks related to data integration [Singh and Jain 2011]. Many DSSPs have been proposed recently addressing a variety of scenarios, e.g., SEMEX [Cai et al. 2005] and iMeMex [Dittrich et al. 2009] on the PIM context; PayGo [Madhavan et al. 2007] focusing on Web-related sources; and a justice-related DSSP[Dijk et al. 2013]. As far as we know, up to date, the proposed DSSPs provide specialized solutions, targeting only specific scenarios [Singh and Jain 2011, Hedeler et al. 2009].

3. LinkedScales: A Multiscale Dataspace Architecture

The goal of LinkedScales is to systematize the dataspace-based integration process in an architecture. It slices integration levels in progressive layers, whose abstraction is inspired by the notion of scales. As an initial effort, LinkedScales strategy focuses on a specific goal on the dataspace scope: to provide a homogeneous view of data, hiding details about heterogeneous and specific formats and schemas. To achieve this goal, the current proposal does not address issues related to access policies, broadcast updates or distributed access management.

LinkedScales is an architecture for systematic and incremental data integration, based on graph transformations, materialized in different scales of abstraction. It aims to support algorithms and common tools for integrating data within the dataspaces. Integration-scales are linked, and data in lower scales are connected to their corresponding representations in higher scales. As discussed in next section, each integration-scale is based on experiences acquired in three previous experiences related to data integration.

Figure 1 shows an overview of the LinkedScales DSSP architecture, presenting, from bottom to top the following scales of abstraction. (i) *Physical Scale*, (ii) *Logical Scale*; (iii) *Description Scale*; and (iv) *Conceptual Scale*.

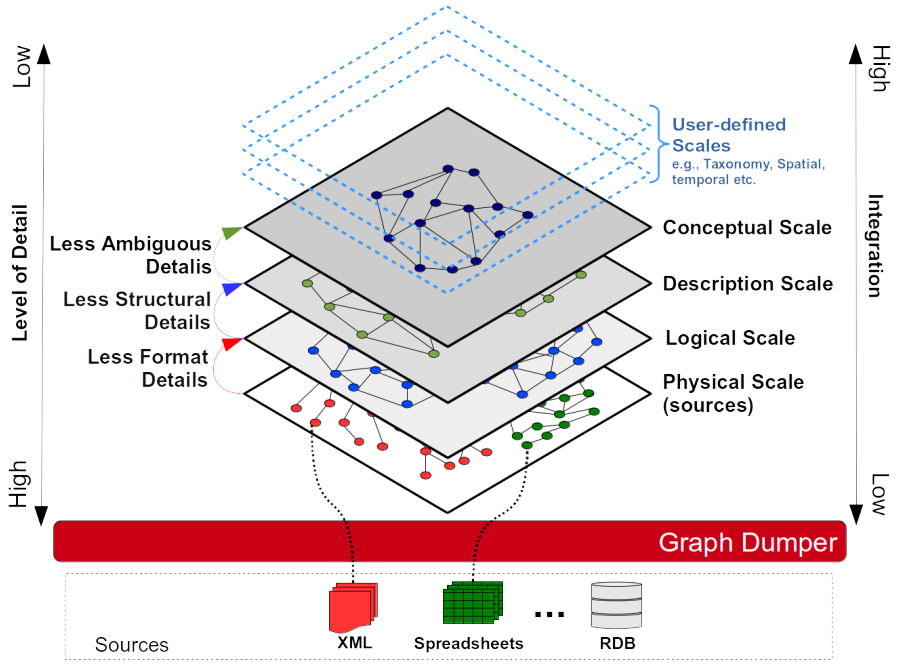

Figure 1. Overview of the LinkedScales architecture.

The lowest part of Figure 1 – the *Graph Dumper* and the *Sources* – represents the different data sources handled by our DSSP in their original format. Even though we are conceiving an architecture that can be extended to any desired format, we are currently focusing on spreadsheets, XML files and textual documents as underlying sources. Data at this level are treated as black-boxes. Therefore, data items inside the sources are still not addressable by links.

The lower scale – the *Physical Scale* – aims at mapping the sources available in the dataspace to a graph inside a graph database. This type of database stores graphs

in their native model and they are optimized to store and handle them. The operations and query languages are tailored for graphs. There are several competing approaches to represent graphs inside the database [Angles 2012, Angles and Gutierrez 2008].

The *Physical Scale* is the lowest-level raw content+format representation of data sources with addressable/linkable component items. It will reflect in a graph, as far as possible, the original structure and content of the original underlying data sources. The role of this scale – in an incremental integration process – concerns making explicit and linkable data within sources. In a dataspace fashion, such effort to make raw content explicit can be improved on demand.

The *Logical Scale* aims at offering a common view to data inside similar or equivalent structural models. Examples of structural models are: table and hierarchical document. In the previous scale, there will be differences in the representation of a table within a PDF, a table from a spreadsheet and a table within a HTML file, since they preserve specificities of their formats. In this (Logical) scale, on the other hand, the three tables should be represented in the same fashion, since they refer to the same structural model. This will lead to a homogeneous approach to process tables, independently of how tables were represented in their original specialized formats. To design the structural models of the Logical Scale we will investigate initiatives such as the OMG's[1] Information Management Metamodel[2] (IMM). IMM addresses the heterogeneity among the models behind Information Management systems, proposing a general interconnected metamodel, aligning several existing metamodels. Figure 2 presents an overview of the current state of the IMM and supported metamodels. For instance, it shows that XML and Relational metamodels can be aligned into a common metamodel.

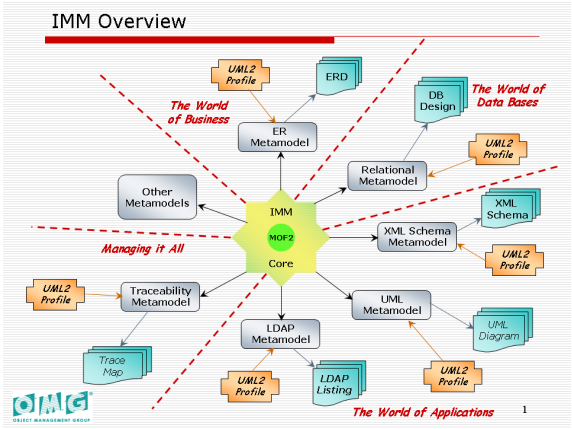

Figure 2. Overview of the current state of the IMM proposal. Source: http://www.omgwiki.org/imm

In the *Description Scale*, the focus is in the content (e.g., labels of tags within a XML or values in spreadsheet cells) and their relationships. Structural information pertaining to specific models – e.g., aggregation nodes of XML – are discarded if they do not affect the semantic interpretation of the data, otherwise, they will be transformed in a relation between nodes following common patterns – for example, cells in the same

[1]Object Management Group – http://www.omg.org
[2]http://www.omgwiki.org/imm

row of a table are usually values for attributes of a given entity. Here, the structures from previous scales will be reflected as RDF triples.

The highest scale of Figure 1 is the *Conceptual Scale*. It unifies in a common semantic framework the data of the lower scale. Algorithms to map content to this scale exploit relationships between nodes of the Description Scale to discover and to make explicit as ontologies the latent semantics in the existing content. As we discuss in next section, it is possible in several scenarios to infer semantic entities – e.g., instances of classes in ontologies – and their properties from the content. We are also considering the existence of predefined ontologies, mapped straight to this scale, which will support the mapping process and will be connected to the inferred entities. Here, algorithms concerning entity linking should be investigated.

4. Previous Work

This proposal was conceived after experiences acquired during three previous research projects. Although with different strategies, they addressed complementary issues concerning data integration. In each project, experiments were conducted in a progressive integration fashion, starting from independent artifacts – represented by proprietary formats, in many cases – going towards the production of connections in lightweight or heavyweight integration approaches. As we will show here, our heavyweight integration here took a different perspective from an upfront one-step integration. It is the end of a chain of integration steps, in which the semantics inferred from the content in the first integration steps influences the following integration steps.

We further detail and discuss the role of each work in the LinkedScales architecture. While [Mota and Medeiros 2013] explores a homogeneous representation model for textual documents independently of their formats, [Bernardo et al. 2013] and [Miranda and Santanchè 2013] focus, respectively, on extracting and recognizing relevant information stored in spreadsheets and XML artifacts, to exploit their latent semantics in integration tasks.

4.1. Homogeneous Model – Universal Lens for Textual Document Formats

One of the key limits to index, handle, integrate and summarize sets of documents is the heterogeneity of their formats. In order to address this problem, we envisaged a "document space" in which several document sources represented in heterogeneous formats are mapped to a homogeneous model we call *Shadow* [Mota and Medeiros 2013].

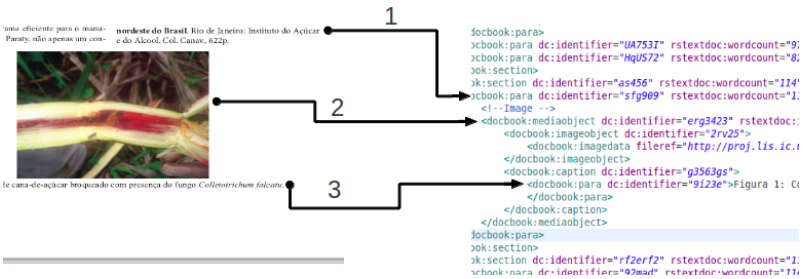

Figure 3. Main idea behind the work [Mota and Medeiros 2013]: A PDF document and its corresponding shadow.

Figure 3 illustrates a typical Shadow (serialized in XML). The content and structure of a document in a specific format (e.g., PDF, ODT, DOC) is extracted and mapped to an open structure – previously defined. The model behind this new structure, which is homogeneous across documents in the space, is a common hierarchical denominator found in most textual documents – e.g., sections, paragraphs, images. In the new document space a shadow represents *format+structure* of a document, decoupled from its specialized format.

Shadows documents are abstractions of documents in specific formats, i.e., they do not represent integrally the information of the source, focusing in the common information that can be extracted according to the context. This abstract homogeneous model allowed us to develop interesting applications in: document content integration and semantic enrichment [Mota et al. 2011]; and searching in a document collection considering structural elements, such as labels of images or references [Mota and Medeiros 2013].

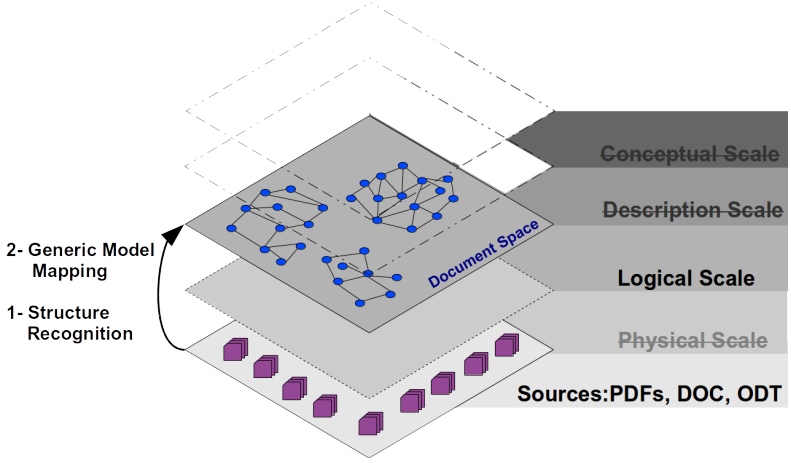

Figure 4. Shadows approach presented in a LinkedScales perspective.

Figure 4 illustrates how this homogeneous view for a document space fits in the LinkedScales architecture. This document space is equivalent to the Logical Scale, restricted to the document context. Different from the LinkedScales approach, Shadows map the documents in their original format straight to the generic model, without an intermediary Physical Scale.

After the Shadows experience we observed three important arguments to represent such intermediary scale: (i) since this scale is not aimed at mapping the resources to a common model, it focus in the specific concern of making explicit and addressable the content; (ii) it preserves the best-effort graph representation of the source, with provenance benefits; (iii) the big effort in the original one-batch-way conversion is factored in smaller steps with intermediary benefits.

In the LinkedScales' *Logical Scale*, the Shadows' document-driven common model will be expanded towards a generic perspective involving a family of models.

4.2. Connecting descriptive XML data – a Linked Biology perspective

[Miranda and Santanchè 2013] studied a particular problem in the biology domain, related to phenotypic descriptions and their relations with phylogenetic trees. Phenotypic

descriptions are a fundamental starting point for several biology tasks, like identification of living beings or phylogenetic tree construction. Tools for this kind of description usually store data in independent files following open standards (e.g., XML). The descriptions are still based on textual sentences in natural language, limiting the support of machines in integration, correlation and comparison operations.

Even though modern phenotype description proposals are based on ontologies, there still are open problems of how to take advantage of the existing patrimony of descriptions. In such scenario, [Miranda and Santanchè 2013] proposes a progressive integration approach based on successive graph transformations, which exploits the existing latent semantics in the descriptions to guide this integration and semantic enrichment.

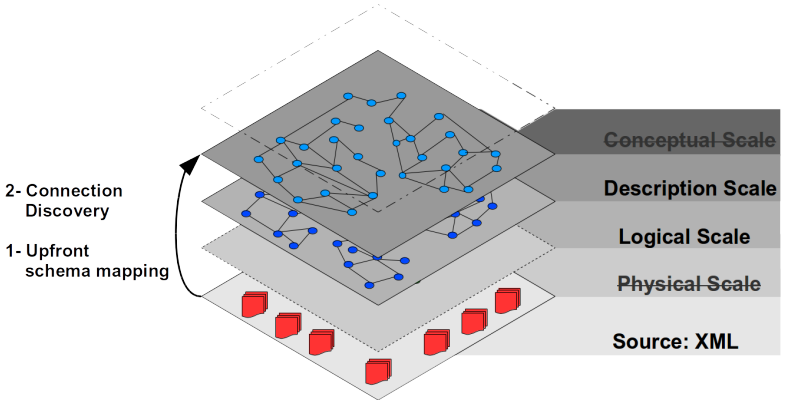

Figure 5. Linked Biology project presented in a LinkedScales perspective.

Since the focus is in the content, this approach departs from a graph-based schema which is a minimal common denominator among the main phenotypic description standards. Operations which analyses the content – discovering hidden relations – drive the integration process. Figure 5 draws the intersection between our architecture and the integration approach proposed by [Miranda and Santanchè 2013]. Data of the original artifacts are mapped straight to the Description Scale, in which structures have a secondary role and the focus is in the content.

In spite of the benefits of the focus in the content, simplifying the structures, this approach loses information which will be relevant for provenance. Moreover, in an interactive integration process, the user can perceive the importance of some information not previously considered in the Description Scale. In this case, since the mapping comes straight from the original sources, it becomes a hard task to update the extraction/mapping algorithms to afford each new requirement. The Physical and Logical Scales simplify this interactive process, since new requirements means updating graph transformations from lower to upper scales.

4.3. Progressively Integrating Biology Spreadsheet Data

Even though spreadsheets play important role as "popular databases", they were designed as self contained units. This characteristic becomes an obstacle when users need to integrate data from several spreadsheets, since the content is strongly coupled to file formats, and schemas are implicit driven to human consumption. In [Bernardo et al. 2013], we decoupled the content from the structure to discover and make explicit the implicit schema embedded in the spreadsheets.

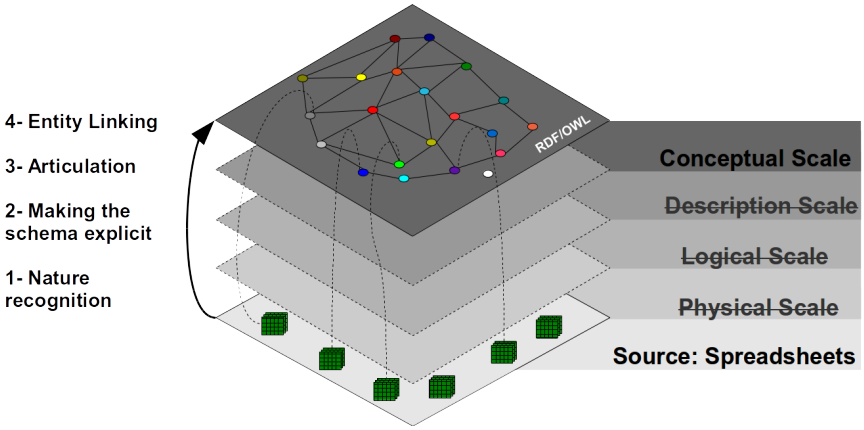

Figure 6. Spreadsheet integration presented in a LinkedScales perspective.

Figure 6 illustrates the [Bernardo et al. 2013] approach in a LinkedScales perspective. The work is divided in four steps, going from the original spreadsheets formats straight to the *Conceptual Scale*. The first step is to recognize the spreadsheet nature. The work assumes that users follow and share domain-specific practices when they are constructing spreadsheets, which result in patterns to build them. Such patterns are exploited in order to capture the nature of the spreadsheet and to infer a conceptual model behind the pattern, which will reflect in an ontology class in the *Conceptual Scale*.

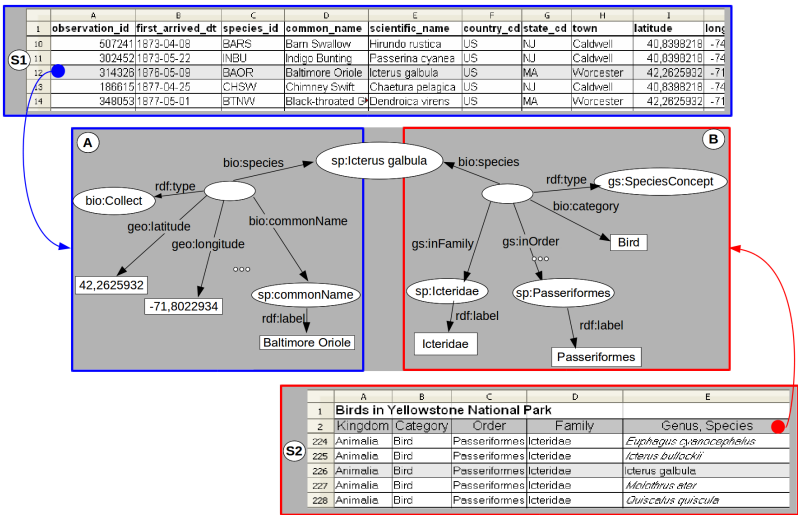

Figure 7. Spreadsheet data articulation via entity recognition.

This work stresses the importance of recognizing data as semantic entities to guide further operations of integration and articulation. Via this strategy, authors are able to transform several spreadsheets into a unified and integrated data repository. Figure 7 shows an example summarizing how they are articulated, starting from the recognition of semantic entities behind implicit schemas. Two different spreadsheets ($S1$ and $S2$) related to the biology domain have their schema recognized and mapped to specific ontology classes – shown in Figure 7 as (A) and (B).

Semantic entities can be properly interpreted, articulated and integrated with other sources – such as DBPedia, GeoSpecies and other open datasets. In an experiment in-

volving more than 11,000 spreadsheets, we showed that it is possible to automatically recognize and merge entities extracted from several spreadsheets.

Figure 8 shows a screencopy of our query and visualization prototype for data[3] extracted from spreadsheets (available in `http://purl.org/biospread/?task=pages/txnavigator`).

This work subsidized our proposal of a Conceptual Scale as the topmost layer of our LinkedScales architecture. Several intermediary steps of transformation from the original datasources towards entities are hidden inside the extraction/mapping program. As in the previous cases, the process can be improved by materializing these intermediate steps in scales of our architecture.

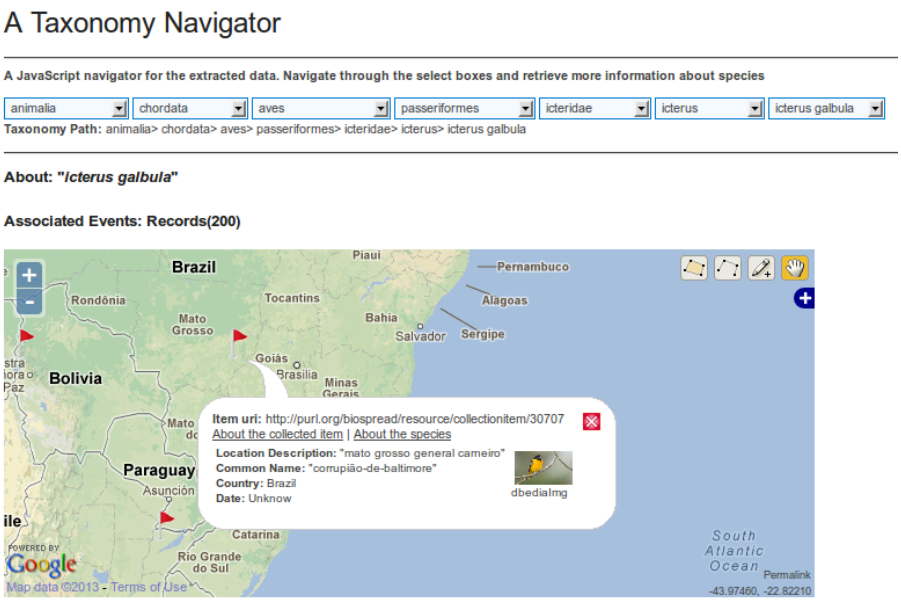

Figure 8. Screencopy of our prototype integrating data of several spreadsheets.

5. Concluding Remarks

This work presented a proposal for a dataspace system architecture based on graphs. It systematizes in layers (scales) progressive integration steps, based in graph transformations. The model is founded in previous work, which explored different aspects of the proposal. LinkedScales is aligned with the modern perspective of treating several heterogeneous datasources as parts of the same dataspace, addressing integration issues in progressive steps, triggered on demand. Although our focus is in the architectural aspects, we are designing a generic architecture able to be extended to several contexts.

Acknowledgments

Work partially financed[4] by CNPq (#141353/2015-5), Microsoft Research FAPESP Virtual Institute (NavScales project), FAPESP/Cepid in Computational Engineering and Sciences (#2013/08293-7), CNPq (MuZOO Project), INCT in Web Science, FAPESP-PRONEX (eScience project), and individual grants from CAPES.

[3]All data is available at our SPARQL endpoint: http://sparql.lis.ic.unicamp.br

[4]The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

[Angles 2012] Angles, R. (2012). A comparison of current graph database models. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*, pages 171–177. IEEE.

[Angles and Gutierrez 2008] Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39.

[Belhajjame et al. 2013] Belhajjame, K., Paton, N. W., Embury, S. M., Fernandes, A. A., and Hedeler, C. (2013). Incrementally improving dataspaces based on user feedback. *Information Systems*, 38(5):656 – 687.

[Bernardo et al. 2013] Bernardo, I. R., Mota, M. S., and Santanchè, A. (2013). Extracting and semantically integrating implicit schemas from multiple spreadsheets of biology based on the recognition of their nature. *Journal of Information and Data Management*, 4(2):104.

[Cai et al. 2005] Cai, Y., Dong, X. L., Halevy, A., Liu, J. M., and Madhavan, J. (2005). Personal information management with semex. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 921–923, New York, NY, USA. ACM.

[Dijk et al. 2013] Dijk, J., Choenni, S., Leertouwer, E., Spruit, M., and Brinkkemper, S. (2013). A data space system for the criminal justice chain. In Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., Leenheer, P., and Dou, D., editors, *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, volume 8185 of *Lecture Notes in Computer Science*, pages 755–763. Springer Berlin Heidelberg.

[Dittrich et al. 2009] Dittrich, J., Salles, M. A. V., and Blunschi, L. (2009). imemex: From search to information integration and back. *IEEE Data Eng. Bull.*, 32(2):28–35.

[Dong and Halevy 2007] Dong, X. and Halevy, A. (2007). Indexing dataspaces. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 43–54, New York, NY, USA. ACM.

[Elsayed and Brezany 2010] Elsayed, I. and Brezany, P. (2010). Towards large-scale scientific dataspaces for e-science applications. In *Database Systems for Advanced Applications*, pages 69–80. Springer.

[Elsayed et al. 2006] Elsayed, I., Brezany, P., and Tjoa, A. (2006). Towards realization of dataspaces. In *Database and Expert Systems Applications, 2006. DEXA '06. 17th International Workshop on*, pages 266–272.

[Franklin et al. 2005] Franklin, M., Halevy, A., and Maier, D. (2005). From databases to dataspaces. *ACM Sigmod Record*, 34(4).

[Halevy et al. 2006a] Halevy, A., Franklin, M., and Maier, D. (2006a). Principles of dataspace systems. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART*, PODS '06, pages 1–9, New York, NY, USA. ACM.

[Halevy et al. 2006b] Halevy, A., Rajaraman, A., and Ordille, J. (2006b). Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment.

[Heath and Bizer 2011] Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*.

[Hedeler et al. 2009] Hedeler, C., Belhajjame, K., Fernandes, A., Embury, S., and Paton, N. (2009). Dimensions of dataspaces. In Sexton, A., editor, *Dataspace: The Final Frontier*, volume 5588 of *Lecture Notes in Computer Science*, pages 55–66. Springer Berlin Heidelberg.

[Hedeler et al. 2010] Hedeler, C., Belhajjame, K., Paton, N., Campi, A., Fernandes, A., and Embury, S. (2010). Chapter 7: Dataspaces. In Ceri, S. and Brambilla, M., editors, *Search Computing*, volume 5950 of *Lecture Notes in Computer Science*, pages 114–134. Springer Berlin Heidelberg.

[Hedeler et al. 2013] Hedeler, C., Fernandes, A., Belhajjame, K., Mao, L., Guo, C., Paton, N., and Embury, S. (2013). A functional model for dataspace management systems. In Catania, B. and Jain, L. C., editors, *Advanced Query Processing*, volume 36 of *Intelligent Systems Reference Library*, pages 305–341. Springer Berlin Heidelberg.

[Hey et al. 2009] Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.

[Kolaitis 2005] Kolaitis, P. G. (2005). Schema mappings, data exchange, and metadata management. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 61–75, New York, NY, USA. ACM.

[Lenzerini 2002] Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA. ACM.

[Madhavan et al. 2007] Madhavan, J., Cohen, S., Dong, X. L., Halevy, A. Y., Jeffery, S. R., Ko, D., and Yu, C. (2007). Web-scale data integration: You can afford to pay as you go. In *CIDR*, pages 342–350. www.cidrdb.org.

[Miranda and Santanchè 2013] Miranda, E. and Santanchè, A. (2013). Unifying phenotypes to support semantic descriptions. *Brazilian Conference on Ontological Research – ONTOBRAS*, pages 1–12.

[Mota and Medeiros 2013] Mota, M. and Medeiros, C. (2013). Introducing shadows: Flexible document representation and annotation on the web. In *Proceedings of Data Engineering Workshops (ICDEW), IEEE 29th International Conference on Data Engineering – ICDE*, pages 13–18.

[Mota et al. 2011] Mota, M. S., Longo, J. S. C., Cugler, D. C., and Medeiros, C. B. (2011). Using linked data to extract geo-knowledge. In *GeoInfo*, pages 111–116.

[Singh and Jain 2011] Singh, M. and Jain, S. (2011). A survey on dataspace. In Wyld, D., Wozniak, M., Chaki, N., Meghanathan, N., and Nagamalai, D., editors, *Advances in Network Security and Applications*, volume 196 of *Communications in Computer and Information Science*, pages 608–621. Springer Berlin Heidelberg.

EDXL-RESCUER ontology: an update based on Faceted Taxonomy approach

Rebeca Barros[12], Pedro Kislansky[1], Laís Salvador[12], Reinaldo Almeida[1], Matthias Breyer[3], Laia Gasparin Pedraza[3] and Vaninha Vieira[12]

[1]Federal University of Bahia

[2]Fraunhofer Project Center for Software and Systems Engineering at UFBA

[3]VOMATEC International GmbH

{rebecasbarros,laisns,vaninha}@dcc.ufba.br,{pedro.kislansky,reifa28}@gmail.com

{matthias.breyer,laia.gasparin}@vomatec-innovations.de

Abstract. *This paper describes an ontology created for the RESCUER[1] (Reliable and Smart Crowdsourcing Solution for Emergency and Crisis Management), a project funded by the European Union and the Brazilian Ministry of Science, Technology and Innovation. RESCUER uses crowdsourcing information for supporting Industrial Parks (InPa) and Security Forces during an emergency situation. The proposal, EDXL-RESCUER ontology, is based on EDXL (Emergency Data Exchange Language), and it aims to be the RESCUER conceptual model related to the coordinating and exchanging of information with legacy systems. The ontology was evaluated with end-users during a workshop and the results show that EDXL-RESCUER is adequate for Emergency and Crisis domain in InPa and Security forces contexts. Specifically, this paper presents an update of EDXL-RESCUER ontology based on a faceted taxonomy approach.*

Resumo. *Este artigo descreve uma ontologia criada para o RESCUER (Reliable and Smart Crowdsourcing Solution for Emergency and Crisis Management), um projeto patrocinado pela União Européia e pelo Ministério de Ciência, Tecnologia e Inovação do Brasil. O RESCUER usa informação do público para apoiar Parques Industriais e Forças de Segurança durante uma emergência. A ontologia proposta, EDXL-RESCUER, é baseada no EDXL (Emergency Data Exchange Language) e pretende ser o modelo conceitual do RESCUER relacionado à coordenação e troca de informação com os sistemas legados. A ontologia foi avaliada com usuários finais durante um workshop, e os resultados mostram que EDXL-RESCUER é adequada para o domínio de Crises e Emergências nos contextos de Parques Industrias e Forças de Segurança. Especificamente, este artigo apresenta uma atualização da EDXL-RESCUER baseada em uma abordagem de taxinomia facetada.*

1. Introduction

Crowdsourcing information (information that comes from different sources: people affected by the incident, eyewitnesses, security forces and others) is becoming widely used as a source of knowledge and solutions for different problems

[1]http://www.rescuer-project.org/

[Beriwal and Cochran 2013][Besaleva et al. 2013][Eccher et al. 2013]. This paper is part of a research project for developing a crowdsourcing solution for emergency management, the RESCUER project [Villela et al. 2013]. RESCUER intends to provide command centers with real-time contextual information related to the emergency through the collection, combination and aggregation of crowdsourcing information, and to support announcements about the emergencies tailored to different audiences (e.g. authorities, affected community and public).

The RESCUER project encompasses four main components as shown in Figure 1:

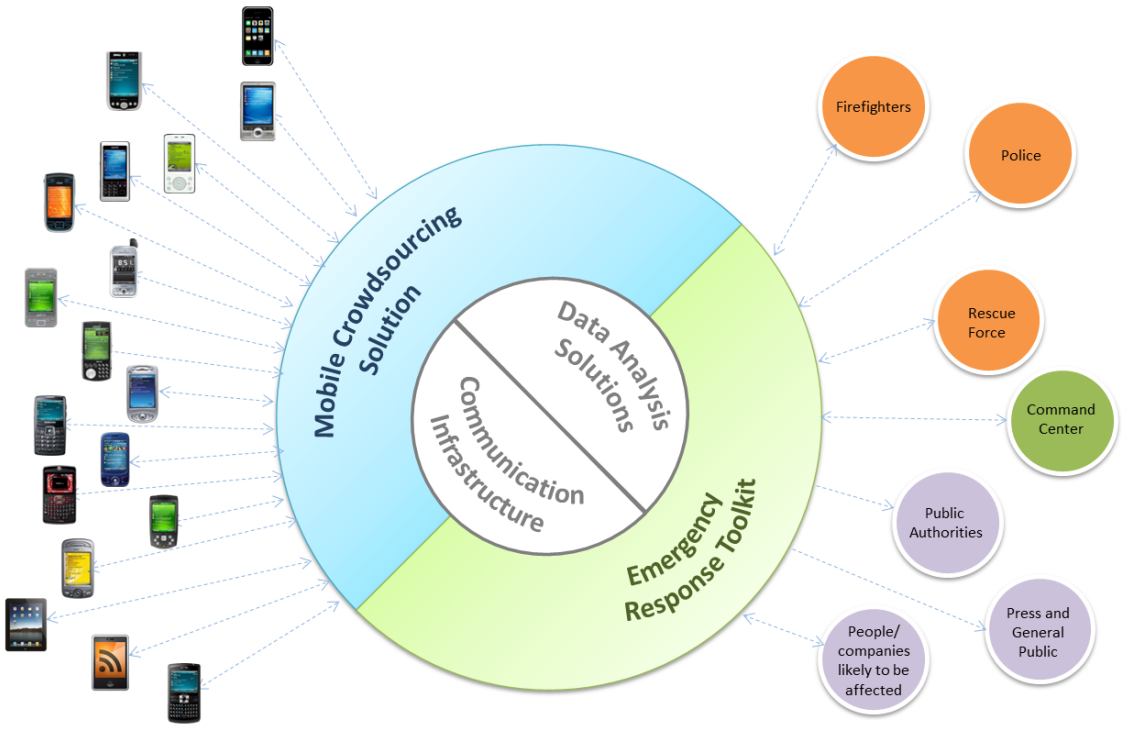

Figure 1. Conceptual model of RESCUER [Villela et al. 2013]

- Mobile Crowdsourcing Solution: support eyewitnesses communication with official first responders (police, firefighters, etc.) and command and control centers. The crowd can send information in text, image and video formats. It comprises a set of mobile applications tailored to different platforms and devices;
- Data Analysis Solutions: composed of the algorithms that will process and filter the data in order to extract the required information;
- Communication Infrastructure: offers the needed equipment in order to allow the information to flow between the stakeholders; and
- Emergency Response Toolkit: a set of solutions to manage the analyzed crowdsourcing information and to present them to the command and control center using adequate visualization metaphors.

The InPa (Industrial Park) Brazilian partner is the COFIC [COFIC 2009] (Industrial Development Committee of Camaçari), which manages security simulations and deals with legal procedures and media. The Security Forces are represented by the CICC

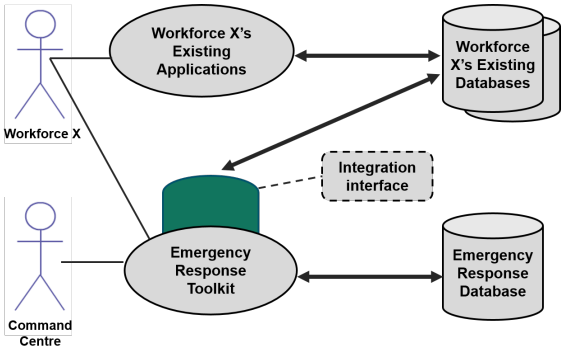

Figure 2. Ontology's role in RESCUER system [Villela et al. 2013]

(Integrated Command and Control Centre) in Brazil and by the FIRESERV[2] in Europe. These partners have contributed to the project with expertise and knowledge on how command and control centers operate in large-scale events, as well as in industrial areas. In this context, interoperability between the RESCUER project and legacy systems' partners is critical for the success of the solution. For the purpose of semantic and seamless integration of legacy systems with RESCUER, the use of ontologies seems to be most suitable, since they offer a basis for a shared and well-formed specification of a particular domain. Thefore, in this proposal, an ontology is presented that will comprise the RESCUER conceptual model related to the coordinating and exchanging of information with legacy systems.

From this perspective, the use of a well-referenced standard by the scientific community, the EDXL [OASIS 2014] – Emergency Data eXchange Language-, as a basis for the new ontology was chosen. EDXL is a common standard that is accepted and used in several applications dealing with disaster management [Genc et al. 2013] [Kilgore et al. 2013]. It is composed of several packages – the current standard version has seven packages, each of which is related to a particular aspect of the emergency domain. A subset of EDXL has been chosen in order to specify the EDXL-RESCUER ontology (details of this process are discussed in section 3). The formalization of the first version of the ontology can be found in our previous paper [Barros et al. 2015]. As an incremental approach is being used, in this paper an update of EDXL-RESCUER ontology based on faceted taxonomy formalization is presented

The Figure 2 shows the ontology-based integration module (green box) in RESCUER architecture, specifically in its interaction with ERTK. This module will have three main parts: 1) the EDXL-RESCUER that works like a global ontology; 2) the database schemas from ERTK and legacy systems involved; 3) mappings between the ontology and the local schemas.

The evaluation of EDXL-RESCUER ontology was performed in two steps:

1. Validation through competency questions - questions that an ontology should be able to answer. This validation is based on a well know method in Ontology Engineering (for further information see the TOronto Virtual Enterprise (TOVE) [Grüninger and Fox 1995] and the METHONTOLOGY

[2]http://www.fireserv.at/

[Fernández-López et al. 1997]).

2. Brainstorming with potential end users for validating the ontology terms. The results show that the EDXL-RESCUER ontology is suitable for specific goals proposed in RESCUER project.

This paper is structured as follows: in the Related Works section, research projects related to emergency, ontology and interoperability are presented. Next, an EDXL-RESCUER ontology and its update based on faceted taxonomy approach is presented. In the Evaluation section the workshop with end-users is described and the results derived are presented; finally, a conclusion of the work done and future developments are presented.

2. Related Works

Ontology has been used on several domains in order to solve interoperability problems, including emergency and crisis domain [Eccher et al. 2013] [Mescherin et al. 2013] [Shah et al. 2013] [Shan et al. 2012] [Xiao et al. 2013].

Based on review of related literature, one project stood out: the DISASTER (Data Interoperability Solution at Stakeholders Emergency Reaction) project [Azcona 2013] [Schutte et al. 2013]. It mainly focuses on Data-Interchange (or more specifically, Data-Artefact-Mapping) on a semantic level. In this project an ontology has been created (EMERGEL) whose main objective was the mapping of different predefined information artifacts, information representations and languages between countries in Europe. In a RESCUER context, the EMERGEL ontology seems to be quite useful as an up-to-date database, if the task of semantically mapping incident information was the objective. For all other aspects needing to be addressed, the interoperability with legacy systems, for instance, EDXL seemed to be more suitable. However, the use of EMERGEL may be investigated in the future for enabling cross-border incidents in Europe.

In addition to the DISASTER project, several works that use ontologies and EDXL in the context of Emergency and Crisis Management were found. Some of these are presented in this section.

The IC.NET (Incident Command NET) is a system that can be used for Emergency Services such as incident representation, triage, and more. It is based on EDXL-DE as a top level loose coupler used for delivery and exposure of operational level Emergency Services / First Responder data [McGarry and Chen 2010].

The TRIDEC[3] project is based on the GITEWS (German Indonesian Tsunami Early Warning System) and the DEWS (Distant Early Warning System). It provides a service platform for both sensor integration and warning dissemination. Warning messages are compiled and transmitted in the OASIS Common Alerting Protocol (EDXL-CAP) together with addressing information defined via the OASIS Emergency Data Exchange Language - Distribution Element (EDXL-DE) [Hammitzsch et al. 2012].

WebPuff is a system sponsored by the U.S. Army CMA(Chemical Material Activity)and developed by IEM, a security consulting firm based in North Carolina's Research Triangle Park. WebPuff provides users at CSEPP (Chemical Stockpile Emergency Preparedness Program) sites with a suite of planning and response tools that are integrated

[3]Project Collaborative, Complex and Critical Decision-Support in Evolving Crises

with a unique chemical dispersion model that provides an advanced level of science on which decisions about public protection can be based.

In order to ensure interoperability with civilian jurisdictions, the system uses the Emergency Data eXchange Language (EDXL) Common Alerting Protocol (CAP) developed by the Organization for the Advancement of Structured Information Standards (OASIS) [Beriwal and Cochran 2013].

The German Research Centre for Geosciences developed a model for integrating the national tsunami warning system on a large scale. They proposed a system based on existing protocols such as EDXL Common Alert Protocol (EDXL-CAP) and the Distribution Element (EDXL-DE) [Lendholt et al. 2012].

3. EDXL-RESCUER Ontology

EDXL is a set of packages of XML-based messaging standards that favor emergency information sharing between organizations and systems. EDXL standardizes messaging formats for communications between these parties. It was developed by OASIS (Organization for the Advancement of Structured Information Standards) [OASIS 2014]

EDXL is a broad enterprise to generate an integrated framework for a wide range of emergency data exchange standards. The EDXL has several packages: EDXL-DE (Distribution Element); EDXL-RM (Resource Messaging); EDXL-SitRep (Situation Reporting); EDXL-HAVE (Hospital Availability Exchange); EDXL-TEP (Tracking of Emergency Patients); EDXL-CAP (Common Alerting Protocol) and EDXL-RIM (Reference Information Model) [OASIS 2014].

An ontology for the semantic integration of data exchange between the RESCUER platform and legacy systems has been defined based on EDXL standards. The current version of EDXL has seven (7) packages and covers a full range of message contexts in an emergency. The extended scope of EDXL has raised several questions, including: (i) Should an ontology be constructed for all packages? (ii) What message contexts are important for RESCUER? (iii) What kind of information will be exchanged with legacy systems?

In order to clear up these doubts, other RESCUER documents related to Requisites and Architecture tasks were analyzed. They were chosen because they provide useful information that can be used in semantic integration of RESCUER with legacy systems. Based upon this study, a list of competency questions can be designed, which serve as a basis for the selection of EDXL packages for RESCUER domain.

Therefore, in order to address these questions, four packages were chosen: EDXL-DE, EDXL-RM, EDXL-SitRep and EDXL-CAP. Four new ontologies were created, one for each chosen package. These were based on ERM and Data Dictionary of their associated standard. These four ontologies comprise the EDXL-RESCUER ontology and the formalization of the first version of them can be found in our previous paper [Barros et al. 2015].

With this first version of the ontology, a validation through competency questions, where each competency question is related with the correspondent ontology elements can be performed, as seen in (Table 1). In this way, the selection of EDXL packages can be validated. This validation also contributes to a first step of evaluation of the ontology.

Table 1. Competency questions X EDXL-RESCUER ontology

Competency Questions	Ontology element correspondent
Where was the incident?	EDXL-RM owl:Class Location EDXL-CAP owl:Class Area
What kind of incident was it?	EDXL-CAP owl:Class Category EDXL-SITREP owl:Class IncidentCause
Which resource (human or material) will be necessary?	EDXL-RM owl:Class RequestResource or another ResourceMessage subclass
When (date and time) did the incident happen?	EDXL-SITREP owl:DataProperty incidentstartdatetime
What is the weather forecast?	EDXL-SITREP owl:DataProperty weatherEffects
How many people have been affected? (deaths, injuries, evacuations)	EDXL-SITREP owl:Class CasualtyandIllnessSummaryReport and related properties
Who reported the incident?	EDXL-DE owl:Class Sender
What kind of message content was sent by the workforces?	EDXL-DE owl:Class ContentDescription

As an incremental approach is being used, in this paper an update of EDXL-RESCUER ontology based on faceted taxonomy formalization as well as its implementation is presented.

3.1. Update of EDXL-RESCUER Ontology

In order to update the EDXL-RESCUER [Barros et al. 2015], we made an in-depth analysis of the data model for the EDXL scheme. During this process, a natural way was to choose Prieto-Diaz proposal [Prieto-Diaz 1987], a technique used for classifying concepts called Faceted Taxonomy. This approach uses a faceted taxonomy with the purpose of improving and reviewing an existing domain ontology. The facets handle three or more dimensions of classification and can be used when it is possible to organize the entities by mutually exclusive and jointly exhaustive categories.

In line with this approach, in [Denton 2003] a method is presented for making a faceted classification using seven steps. These steps adapted for EDXL-RESCUER ontology update are shown below:

a) Domain collection: we used the EDXL Documentation;
b) Entity listing: we listed all entities found;
c) Facet creation: we arranged all entities that resembled under a main entity, the facet (main entity was chosen to represent a domain segment EDXL);
d) Facet arrangement: we made sure that the entities resembled to the associated facets, reorganizing them when appropriate, (the checks were made through the EDXL documentation, which contains the description and data model for the entities).
e) the citation order and f) classification – phases that refer to how the taxonomy would be implemented. In our case, the goal was the creation of an ontology, then

we defined what every element under a facet and the facet itself would be in an OWL ontology, i.e. what is a class, sub-class, object property, and data property.
g) The last phase included revision, testing, and maintenance: the result of this phase is EDXL-Rescuer v2.

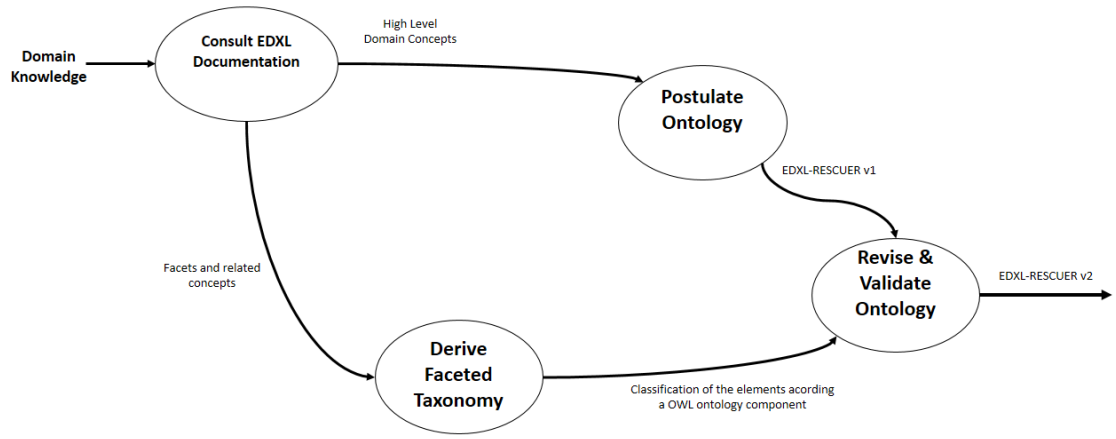

Figure 3. Review and building process of ontologies - Based on [Prieto-Díaz 2003]

Figure 3 summarizes the entire process of the EDXL-RESCUER update. The first version of the ontologies that composed the EDXL-RESCUER relied on EDXL documentation and the ERM models available there. Hence, a faceted taxonomy based on the same documentation, which allowed one to better detail the domain of each chosen pattern was created. Moreover, we were able: (i) to determine the main concepts with higher precision; and (ii) to use the results for reviewing and revalidate the ontologies created at the first iteration.

For instance, the concepts Severity, Urgency and Certainty found in EDXL-CAP. After the procedure previously mentioned (the concepts reviewing), those concepts became classes instead of DataProperties. Those classes received sub-classes with the ability to have different values according to the EDXL documentation as can be seen in Figure 4.

Another improvement from the previous version is that we were able to reuse common concepts among more than one type of EDXL pattern. Hence, Severity, Urgency and Certainty, which EDXL-SitRep also employs, they are imported concepts from EDXL-CAP; therefore the URI is the same as found on the original ontology.

The approach based on faceted taxonomy seemed to be adequate, considering that this technique for classifying concepts is characterized by randomly choosing the terms that represent concepts within a domain (facets). Furthermore, it chooses the relationship between other domain terms and the terms previously chosen, creating categories (each of which is related with a facet). Finally, the faceted approach selects the terms and the relationship between them within the same category or between categories [Dahlberg 1978]

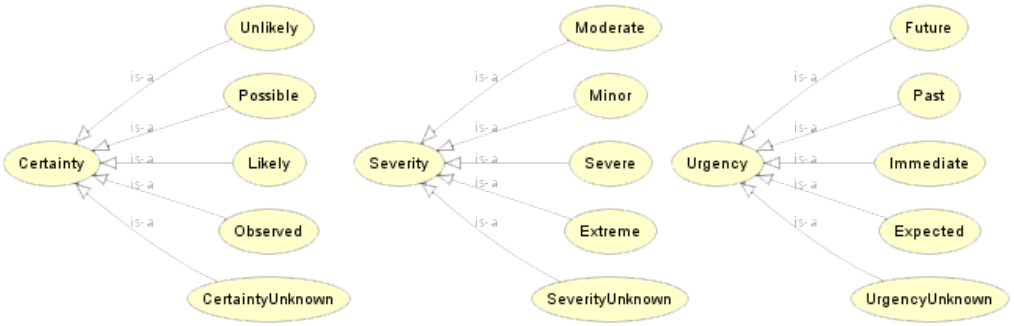

Figure 4. Concepts Urgency, Severity and Certainty - Partial Taxonomy of EDXL-CAP

[Prieto-Díaz 1990]. Additionally, a faceted approach relies not on the breakdown of a universe of knowledge, but on building up or synthesizing from the subject statements of particular documents and that facet can be constructed as perspectives, viewpoints, or dimensions of a particular domain [Prieto-Díaz 2003].

Table 2. Relationship definitions (EDXL-CAP)

Concept1	Relationship	Concept2	Restriction
AlertMessage	hasIncidentRelated	Incident	some
AlertMessage	hasInfo	Info	Min 0
AlertMessage	hasMsgType	MsgType	Max 1
AlertMessage	hasScope	Scope	Max 1
AlertMessage	hasStatus	Status	Max 1
AlertMessage	hasSender	Sender	Max 1
Info	hasArea	Area	some
Info	hasCategory	Category	Max 1
Info	hasResource	Resource	some
Info	hasResponseType	ResponseType	Max 1
Info	hasCertainty	Certainty	Max 1
Info	hasSeverity	Severity	Max 1
Info	hasUrgency	Urgency	Max 1

3.2. Implementation

Due to space limitation, only part of the EDXL-RESCUER ontology is shown. The concepts that make up the EDXL-CAP Ontology and their definitions are:

- AlertMessage: Refers to all component parts of the alert message.
- Info: Refers to all component parts of the info sub-element of the alert message.
- Resource: Necessary element to deal with an emergency. A Resource contains information about its Identity, Description and Status.
- Incident: Term referring to occurrences of any scale that may require some form of Emergency Response and Management, and that requires tracking and information exchange.
- ResponseType: Refers to the type of action recommended for the target audience.

- Area: Refers to all component parts of the area sub element of the info sub element of the alert message.
- Category: Refers to the category of the subject event of the alert message
- MsgType: Refers to the nature of the alert message.
- Status: Refers to the appropriate handling of the alert message.
- Scope: Refers to the intended distribution of the alert message.
- Sender: The originator of an alert.
- Certainty: The certainty of the subject event of the alert message
- Severity: The severity of the subject incident or event.
- Urgency: The urgency of the subject event of the alert message

Table 2 presents the definition of their relationships. The following semantics are used:

Zero or more objects of `<Concept1>` `<Relationship>` with `<Restriction>` objects of `<Concept2>`.

Where `<Restriction>` can be `some, all, Max 1, Min 0, Exactly 1`. `Min 0` is the default value.

Some axioms have also been defined, for instance: (i) Private, Public and Restricted - subclasses of Scope – are disjoint concepts; (ii) Actual, Draft, Exercise, System and Test – subclasses of Status - are disjoint concepts too.

4. Evaluation

The evaluation occurred during the RESCUER Brazilian Consortium Meeting on July 21-23, 2014 and had the goal of validating the terms with potential RESCUER users in Brazilian side. Next, the Goals, Method and Results of this evaluation will be presented.

4.1. Goals

- To present some ontology terms to the stakeholders - terms which were chosen because they represent the main classes of the selected EDXL packages and were the most controversial for both industrial parks (InPa) and large-scale events (LSE);
- To match those terms with the vocabulary the stakeholders use on a daily basis in order to extract synonyms and verify differences, if differences exist, between InPa and LSE.

4.2. Method

The "brainstorm technique" was used in order to capture stakeholder feedback concerning the ontology terms.

The stakeholders were divided into two groups;

- Industrial parks (COFIC)
- Large-scale events (CICC)

During this session, the EDXL concepts were shown to the experts and they tried to find synonyms or correlated terms used in their contexts. At the end of the session, there was an open discussion about the findings related to main concepts of EDXL-RESCUER ontology.

4.3. Results

Based on the activity conducted with the stakeholders, it can be deduced:

- The concepts related to EDXL-SitRep package, in the COFIC context, were suitable;
- Some concepts, for instance the term "incident", had minor variations between the two groups;
- Almost all EDXL terms had related instances or synonyms according to this activity.
- The exception was the term "Jurisdiction", which did not have an instance or a synonym for COFIC. However, at CICC, was found a related instance.
- Some collected terms can be used as instances for populating the EDXL-RESCUER ontology in the future.

This activity raised some important conclusions:

- The necessity of validating all concepts with Brazilian stakeholders;
- A deep investigation of the differences between industrial parks and large-scale events in Brazil; and
- The need to replicate this activity in the European scenario

It is important to note that the differences between the scenarios (COFIC and CICC) emphasize the need for an Interlingua and the relevance of this proposal - EDXL-RESCUER as a common basis for communication.

5. Conclusion

This paper discuss the conceptual model for semantic integration – EDXL-RESCUER ontology. It aims to integrate, semantically, the RESCUER system with legacy systems. In particular, this paper presents an updated version of the ontologies that composed EDXL-RESCUER based on a faceted taxonomy approach. This approach relied on a bottom-up analysis of the EDXL documentation in order to synthesizing the subject statements of these documents. It is important to note that the construction of facets provides different perspectives and views of the domain. In this way, we were able to review our first version of EDXL-RESCUER ontology and adjust its concepts and relationships.

Moreover, in regards to the evaluation, the legacy systems information and data are still missing, as well as the data from RESCUER base. After populating the EDXL-RESCUER ontology, we are going to validate it using reasoning algorithms and queries. Another step is to implement the ontology-based integration module between RESCUER and legacy systems.

Some further investigations will be carried out as well: (i) the use of LOD (Linked Open Data) in this context; (ii) the use of the EMERGEL-knowledge base as an additional controlled vocabulary or just as a synonym-base.

6. Acknowledgements

This research is part of the RESCUER (Reliable and Smart Crowdsourcing Solution for Emergency and Crisis Management) project funded by European Union under grant reference 614154 and by CNPq under grant reference 490084/2013-3. This work also was partially supported by the National Institute of Science and Technology for Software Engineering (INES), funded by CNPq, grant 573964/2008-4.

References

Azcona, E. R. (2013). Disaster data interoperability solution at stakeholders emergency reaction. http://disaster-fp7.eu/sites/default/files/D3.22.pdf.

Barros, R., Kislansky, P., Salvador, L., Almeida, R., Breyer, M., and Gasparin, L. (2015). Edxl-rescuer ontology: Conceptual model for semantic integration. In *Proceedings of the 12th International ISCRAM Conference*.

Beriwal, M. and Cochran, B. (2013). Protecting communities from chemical warfare agents. In *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*, pages 418–422. IEEE.

Besaleva, L., Weaver, A. C., et al. (2013). Applications of social networks and crowdsourcing for disaster management improvement. In *Social Computing (SocialCom), 2013 International Conference on*, pages 213–219. IEEE.

COFIC (2009). Cofic polo - comitê de fomento industrial de camaçari. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency.

Dahlberg, I. (1978). Teoria do conceito. *Ciência da informação*, 7(2).

Denton, W. (2003). How to make a faceted classification and put it on the web. https://www.miskatonic.org/library/facet-web-howto.html.

Eccher, C., Scipioni, A., Miller, A. A., Ferro, A., and Pisanelli, D. M. (2013). An ontology of cancer therapies supporting interoperability and data consistency in eprs. *Computers in biology and medicine*, 43(7):822–832.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

Genc, Z., Heidari, F., Oey, M. A., van Splunter, S., and Brazier, F. M. (2013). Agent-based information infrastructure for disaster management. In *Intelligent Systems for Crisis Management*, pages 349–355. Springer.

Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies.

Hammitzsch, M., Reißland, S., and Lendholt, M. (2012). A walk through tridec's intermediate tsunami early warning system. In *EGU General Assembly Conference Abstracts*, volume 14, page 12250.

Kilgore, R., Godwin, A., Hogan, C., et al. (2013). A precision information environment (pie) for emergency responders: Providing collaborative manipulation, role-tailored visualization, and integrated access to heterogeneous data. In *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*, pages 766–771. IEEE.

Lendholt, M., Esbri, M., and Hammitzsch, M. (2012). Interlinking national tsunami early warning systems towards ocean-wide system-of-systems networks. In *Proceedings of the 9th international ISCRAM conference, Vancouver, Canada*.

McGarry, D. P. and Chen, C. R. (2010). Ic. net—incident command "net": A system using edxl-de for intelligent message routing. In *Technologies for Homeland Security (HST), 2010 IEEE International Conference on*, pages 197–203. IEEE.

Mescherin, S., Kirillov, I., Klimenko, S., et al. (2013). Ontology of emergency shared situation awareness and crisis interoperability. In *Cyberworlds (CW), 2013 International Conference on*, pages 159–162. IEEE.

OASIS (2014). Oasis emergency management tc. `https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency`.

Prieto-Diaz, R. (1987). Domain analysis for reusability. In *Proceedings of COMPSAC*, volume 87, pages 23–29.

Prieto-Díaz, R. (1990). Domain analysis: An introduction. *ACM SIGSOFT Software Engineering Notes*, 15(2):47–54.

Prieto-Díaz, R. (2003). A faceted approach to building ontologies. In *Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference on*, pages 458–465. IEEE.

Schutte, F., Rubén, C., and Emilio, R. (2013). Solving interoperability issues in cross border. In *Proceedings of the 10th International ISCRAM Conference*, pages 238–241.

Shah, T., Rabhi, F., and Ray, P. K. (2013). Oshco: A cross-domain ontology for semantic interoperability across medical and oral health domains. In *e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*, pages 460–464. IEEE.

Shan, G., Rui, G., Wenjun, W., and Xiankun, Z. (2012). Research on ontology-based emergency situation data integration and sharing. *Journal of Convergence Information Technology*, 7(9).

Villela, K., Vieira, V., Mendonça, M., Torres, J., and Graffy, S. (2013). Rescuer-dow - reliable and smart crowdsourcing solution for emergency and crisis management.

Xiao, H., Qiu, T., and Zhou, P. (2013). Integration of heterogeneous agriculture information system based on interoperation of domain ontology. In *Agro-Geoinformatics (Agro-Geoinformatics), 2013 Second International Conference on*, pages 476–480. IEEE.

Measurement Ontology Pattern Language Applied to Network Performance Measurement

Raphaela P. Nunes[1], Adriana S. Vivacqua[1], Maria Luiza M. Campos[1], Ana Carolina Almeida[2]

[1]Programa de Pós-Graduação em Informática - Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.530 – 21941-590 – Rio de Janeiro – RJ - Brasil

[2]Instituto de Matemática e Estatística - Departamento de Informática e Ciências da Computação - Universidade do Estado do Rio de Janeiro (UERJ)
Rua São Francisco Xavier, 524 - 6º andar - 20550-013 - Rio de Janeiro - RJ - Brazil

{mluiza,raphaela.nunes}@ppgi.ufrj.br, avivacqua@dcc.ufrj.br, ana.almeida@ime.uerj.br

Abstract. *The Measurement Ontology Pattern Language (M-OPL) addresses the measurement core conceptualization according to an Ontology Pattern Language (OPL). An OPL provides holistic support for solving ontology development problems for a particular field and guiding the development of ontologies. This paper presents the application of M-OPL in a specific domain, network performance measurement. As a result of this application, a discussion of the use of M-OPL is presented together with some suggestions of extensions to contemplate the peculiarities of this domain.*

1. Introduction

Measurement is a very important discipline in several domains, since it provides useful information for getting conclusions and making decisions [BARCELLOS *et al*., 2014]. Measurement is the process of assigning numbers or symbols to properties of real-world entities, according to widely-defined rules, in order to describe them [FINKELSTEIN; LEANING, 1984]. It can also be understood as a process that involves a set of actions in order to characterize entities assigning values to their properties [BARCELLOS *et al*., 2014].

When analyzing different areas where measurements can be applied, it is possible to identify some particular concepts related to the knowledge treated on each specific area. However, it is also possible to identify some core concepts that are independent of the application domain. In order to homogeneously represent these core concepts across different domains, avoiding inconsistencies and ambiguities, it is important to use a common terminology shared by the domains. Currently, core ontologies have been used to promote this common conceptualization.

The Measurement Ontology Pattern Language (M-OPL) [BARCELLOS *et al*., 2014] addresses the main conceptualization associated to measurements in general, organized according to an Ontology Pattern Language (OPL). An OPL [FALBO *et al.*, 2013] corresponds to a network of interconnected ontology modeling patterns that

provides holistic support for solving ontology development problems for a particular field.

In this work, we describe the application of M-OPL to the scenario of measurements associated to performance monitoring of Internet links. The objective is to discuss the scope and usage of M-OPL to generate a new version of the original ontology developed in the context of the Pinger-LOD Project [SOUZA *et al.*, 2014]. By aligning it to the core modeling patterns proposed for measurements, we aim to decrease the possibility of inconsistencies and ambiguities on the ontology and facilitate future publication and linkage of data with other related data sources in the Web.

During the application of M-OPL we are also concerned to represent multidimensional aspects of measures, to support the representation of different perspectives of a measurement. As in other domains, a proposed extension to an existing OPL may derive, in the future, new modeling patterns to be incorporated in a new version of the OPL, in this case, an extended version of M-OPL.

The remainder of this paper is organized as follows: Section 2 presents a brief description of M-OPL. In Section 3, we describe the application process of M-OPL to the network measurement domain, making firstly a brief explanation of the current structure of PingER ontology and how it was derived. In Section 4, related works and further discussions are presented. Finally, in Section 5, we conclude and list some future work.

2. Measurement Ontology Pattern Language (M-OPL)

A core ontology provides a precise definition of the structural knowledge in a specific field that spans across several application domains in that field [SCHERP *et al.*, 2012]. Ontology Pattern Languages (OPL) have been proposed and used to organize core ontologies facilitating their reuse and extension [FALBO *et al.*, 2013]. An OPL [FALBO *et al.*, 2013] provides a set of interconnected ontology modeling patterns and a process that describes how to combine them to build an ontology applied to a specific domain.

The M-OPL addresses the core conceptualization for measurements and their characterization. M-OPL includes six patterns, defined according to the Unified Foundational Ontology (UFO) [GUIZZARDI, 2005] and covering six measurements aspects: Measurement Entities, which include patterns related to the entities and their properties that can be measured; Measures, which deal with the definition of measures and classify them according to their dependence on other measures; Measurement Units & Scales, which concerns the scales related to measures and the measurements units used to partition the scales; Measurement Procedures, which deals with procedures required to collect data for measures; Measurement Planning, which addresses the goals that drive measurement and the measures used to verify goals achievement; and Measurement & Analysis, which concerns data collection and analysis.

In the application of M-OPL discussed in this article, we will be particularly interested in applying the Measures aspects, in order to define the measures associated to the network links measurement domain and characterize the main entity to be measured (Measurable Entity) in the field, which, in this case, is the Internet Link. For

the purpose of this paper, we are not including details on measurement procedures or data collection.

3. M-OPL Application to Network Performance Measurement

In this section, we discuss how M-OPL was used to derive a new version of an ontology for the conceptualization of measurements for network links performance. The original ontology was developed in the context of the PingER (Ping End-to-end Reporting) project, which is conducted by the Network and Telecommunications Department at the SLAC[1] National Accelerator Laboratory, in Stanford University, USA. The project manages data about the quality of Internet links from 1998 to the present day, on an hourly and daily basis, comprising 16 different metrics collected by 80 monitor nodes over 800 monitored nodes (more than 8000 pairs of nodes), in more than 160 countries [COTTRELL, 2001]. Each measurement is basically defined by a ping sent from a monitor node to a monitored node at any given time, and related to a specific network metric, considering data packets sizes of 100 and 1000 bytes.

3.1. Original PingER Ontology

The original PingER ontology [SOUZA *et al*., 2014] was developed to serve as a reference vocabulary and structure to represent and annotate PingER data as RDF triples for a linked data publishing and querying application.

The PingER ontology is an adaptation of the MOMENT (Monitoring and Measurement in the Next Generation Technologies) ontology [SALVADOR. *et al.*, 2010; RAO, 2010], a core ontology which conceptualizes the networking performance measurements domain.

The MOMENT ontology is complex and generic in the way it contemplates the main characteristics referring to network measurement. This generality of the ontology enables it to be adapted to many different network measurement scenarios, including the PingER domain. However, since the ontology is so generic, the ontology fails in representing PingER reality. Additionally, the ontology does not aim to minimize the number of triples generated, which make it harder to process a large amount of data. Thus, it was decided not to reuse the ontology exactly as it is, but, instead, to reuse its concepts and ideas as basis to build an ontology more specialized for the PingER domain, which could better support data analytical processing. The current version of Pinger ontology has been implemented and used to publish PingER data, that can be accessed from a SPARQL endpoint[2].

Figure 1 shows an overview of the generated model. In the center of the ontology is the main superclass, which is the *Measurement* class, representing the process of acquiring measures. *Measurement* relates to the following classes, in order to qualify the measurement: *Metric*, through *measuresMetric* relation to specify which network metric is being measured; *MeasurementParameters* which can be specialized in *PacketSize*, through *hasMeasurementParameters* relation, to specify the measurement attributes; *DateTime* through *hasDateTime* relation, to specify the time interval in which the

[1] https://www6.slac.stanford.edu/

[2] http://pingerlod.slac.stanford.edu/sparql

measurement was made; *SourceDestinationNodes,* which represents the Internet Links and is related to two types of *Network Nodes*, the one which performs the role of monitor node, sending the ping signal, and the other which performs the role of monitored node, receiving the ping. The relation is made through *PingER-ont:hasSourceNode* and *PingER-ont:hasDestinationNode* relations, respectively.

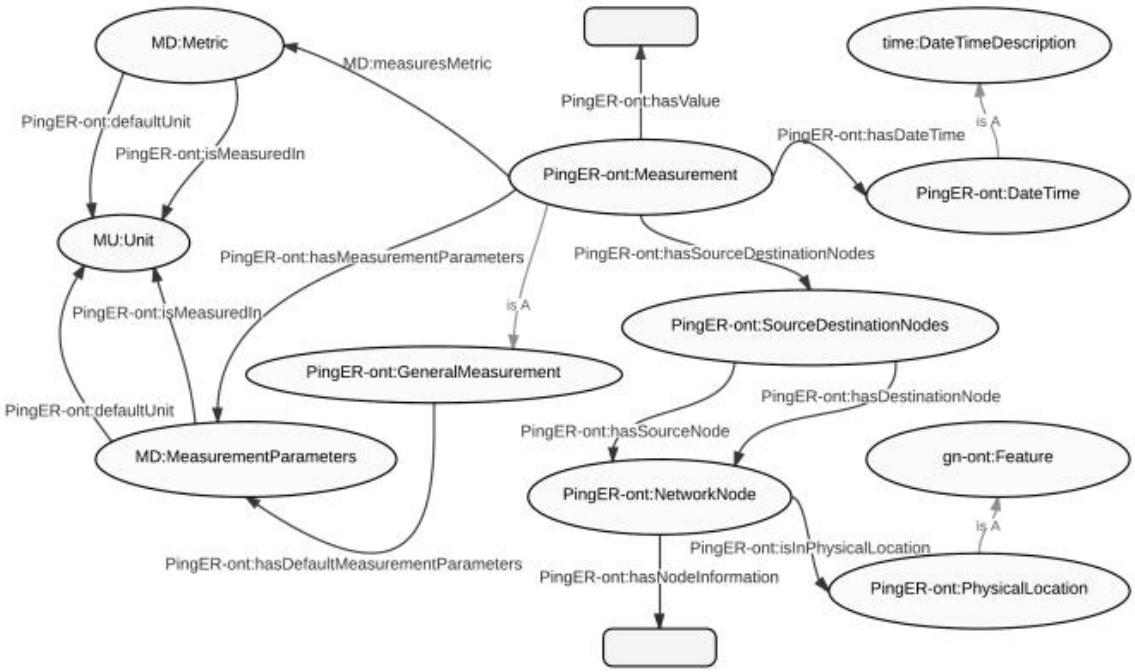

Figure 1. PingER Ontology (SOUZA *et al*., 2014)

To define the parameters of time (when the measure was taken) and space (where the network nodes - NetworkNode - are located), concepts extracted from Time [HOBBS & PAN, 2006] and Geonames [VATANT & WICK, 2012] ontologies were used.

3.2. M-OPL Application to PingER Network Performance Measurement

In order to apply M-OPL to the network performance measurement domain, we used the patterns depicted in gray in Figure 2, applied in the order indicated by the darker lines. The process was defined in the paper that originally presented M-OPL [BARCELLOS *et al*., 2014]. Figure 3 shows a fragment of the resulting ontology.

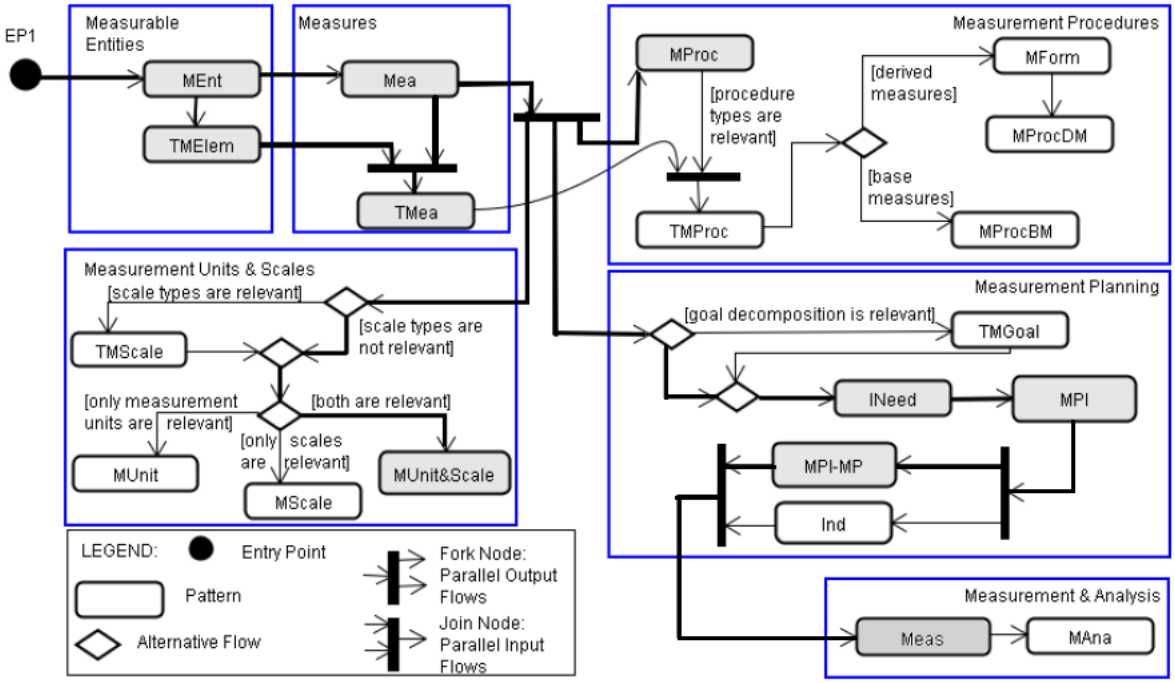

Figure 2. Application order of modeling patterns

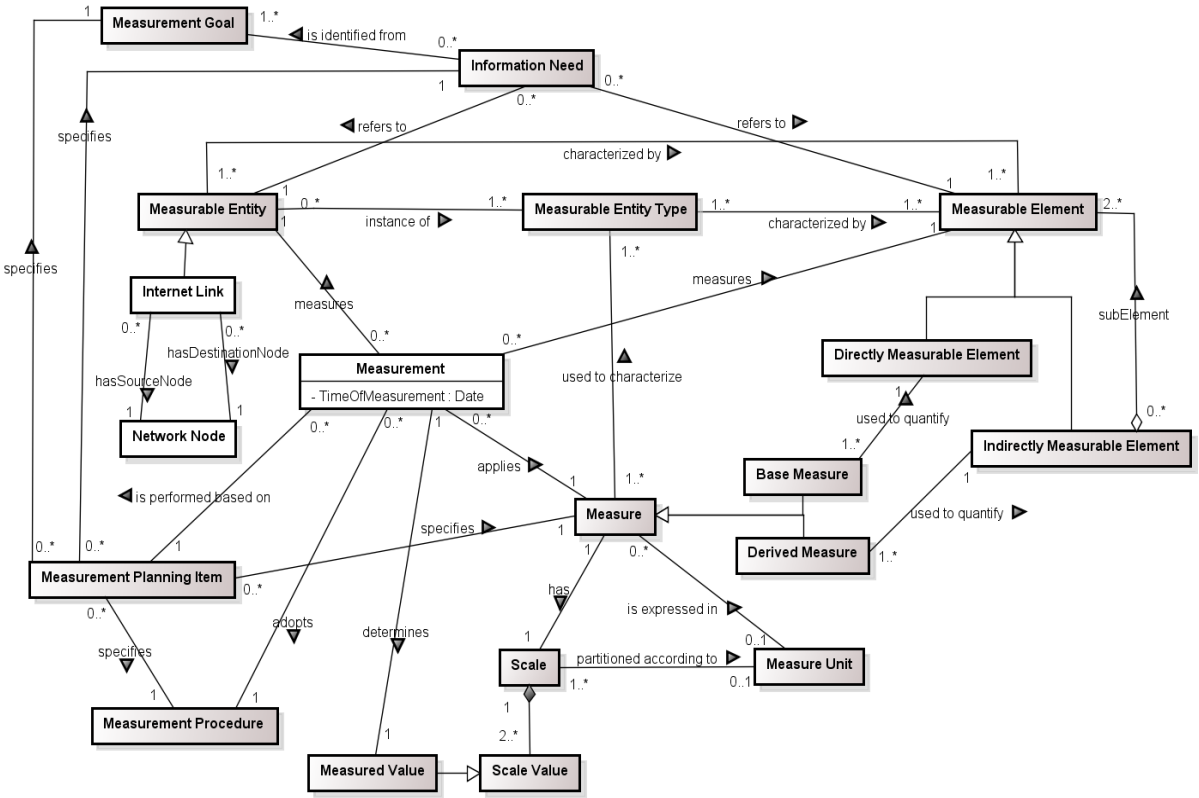

Figure 3. Derived Ontology

The first pattern applied was **MEnt** (Measurable Entity), which has been extended to consider the type of measurable entity relevant to the domain, an I*nternet link*. This is the current Measurable Entity Type being monitored by the PingER project, but others could be considered and then included.

After using the pattern **MEnt**, two patterns were applied: **TMElem** (Types of Measurable Elements) and **Mea** (Measures). In pattern **TMElem**, we could identify the Measurable Elements considered by the PingER ontology structure and characterize them as Directly Measurable Elements (elements that do not depend on others to be measured) or Indirectly Measurable Elements (elements that depend on other sub-elements to be measured).

Examples of Directly Measurable Elements are Duplicate Packets and Packet Loss, as they result from counting the associated events. Indirectly Measurable Elements depend on sub-elements to be measured, and in PingER case they include Directivity, Minimum Round Trip Delay, Conditional Loss Probability, Mean Opinion Score and Average Round Trip Time, among others. Round Trip Time or RTT, for example, is related to the distance between the nodes plus the delay at each hop along the path between them.

The **Mea** pattern was used as it was defined in M-OPL. In this pattern, a Measure quantifies a Measurable Element, characterizing a Measurable Entity Type. Hence, the measure *number of packets* quantifies the measurable element *packet loss* that characterizes the measurable entity of type *Internet Link*. But to better define an Internet Link it was necessary to extend the M-OPL, adding the concept of *NetworkNode*, which was related to *Internet Link* through *hasSourceNode* and *hasDestinationNode* relations, employed according to the role that the node is performing during measurement.

As the pattern **TMElem** was used, the pattern **TMea** (Type of Measures) also had to be used to characterize a Measure as a Base or Derived Measure, which serves to quantify Directly and Indirectly Measurable Elements, respectively. This pattern was also applied exactly as it was defined in M-OPL.

After using the **Mea** pattern, three paths were followed in parallel. The first led to the Measurement Units & Scales group, the second to the Measurement Procedures group and the third to the Measurement Planning group. In the Measurement Units & Scales group, as it is important for the domain in order to model the units and scales of measures, the pattern **MUnit&Scale** was used, as it was defined in M-OPL.

In the Measurement Procedures Group, as it is not important for the domain to detail the data collection procedures according to the different types of measures, only was used, in this group, the pattern **MProc**, as it was defined in M-OPL.

In the Measurement Planning Group, the first pattern used was **INeed**. For example, in our case, *Know the variability of service* could be considered as an instance of Information Need. It could be used to indicate the achievement of the Measurement Goal, which was defined in PingER domain as *Check the network quality*. Although not represented in the fragment of Figure 4, measurement goals may be composed or

simple. In this case, *Check the network transfer capacity* could be a sub-goal of the composed measurement goal *Check the network quality*.

The next pattern used after **INeed** was **MPI-MP** (Measurement Planning Item – Measurement Procedure), which was applied in the same way that it was defined in the M-OPL. Finally, after addressing a Measurement Planning Item, the **Meas** (Measurement) pattern of Measurement & Analysis group was applied. In the **Meas** pattern, Measurement is executed based on a Measurement Planning Item and adopting a certain Measurement Procedure. It measures a Measurable Element of a Measurable Entity applying a Measure. The result is a Measured Value, which refers to a value of a measure scale.

To be able to represent temporal aspects in **Meas** pattern, *TimeOfMeasurement* was added as property of the relator Measurement (actually a property of the Event giving rise to the Relator). In UFO, Relators are derived from Events, which are temporal based constructs.

Making a brief comparison between the ontology derived from the application of M-OPL and the original PingER ontology, it is possible to note that the original version of PingER ontology is more focused on treating the particular domain concepts, representing only partially the semantics of measurements. M-OPL includes a general knowledge about measurements, applicable to different situations. By applying the M-OPL, these generic classes can be specialized according to the situation being considered. For example, in our scenario, the main focus of network measurements was performance evaluation, considered as quality measures associated to the network (using the Ping procedure as in the SLAC laboratory). However, by specializing M-OPL classes, we can use further grouping of goals and measures, and represent network evaluations other than related to performance/quality, like network reliability measures (which would include Medium Time Between Failures – MTBF, Gracefull Degradation, Recovery Time after Failures, Medium Time to Repair – MTTR, among others) [BALTRUNAS et al, 2014].

Considering the main patterns proposed in M-OPL and comparing with the classes in original structure of PingER ontology, it is possible to note that some of these patterns are already somehow represented in the original ontology structure. The Metric class is similar to the Types of Measurable Elements pattern (**TMElem**), since it is possible to represent the Measurable Elements through this class. But is not possible to distinguish the Measurable Elements which depend or not on others to be measured through this class, then it was necessary to add these new concepts in the ontology, by adding the Directly Measurable and Indirectly Measurable Elements.

The Unit class is similar to the Measurement Units and Scales pattern (**MUnit&Scale**), since it is possible to represent the units in which measures are expressed through this class. But it is not possible to represent in the original ontology the scales for measures which are partitioned according to the units, so it was necessary to create a new class to represent the Scale element and relate it with the Measure Unit class.

The Measurement class is similar to the Measurement pattern (**Meas**), since it functions, like in M-OPL, as a Relator, connecting the classes involved in the

measurement process. However, the measurement process of the original ontology does not consider a Measurement Planning Item neither a Measurement Procedure, so it was necessary to create new classes to represent these elements and relate them with the Measurement class.

By using a generic conceptualization, the derived ontology allows interoperation with other complementary domains, which is particularly interesting for Semantic Web applications and publication of LOD.

4. Related Work

In the literature, two works were found applying the M-OPL. In the original proposal of M-OPL [BARCELLOS *et al*., 2014], it was used to build a Software Measurement Ontology (SMO), with a very straightforward application of the patterns. In [FRAUCHES, 2014], knowledge about the measuring process and the vocabulary adopted in the process described in M-OPL were used, as a basis for defining an approach for obtaining indicators from open data. The approach proposes a set of activities that must be performed from established measurement goals, to organize the data from an open database and to extract indicators that provide useful information for decision making.

Applications of OPL in different domains have also been presented, most of them confirming the possibility of reuse of the proposed patterns, and the usefulness of the accompanying guiding process for their application. There is, though, still a lack of cases where the derived domain ontologies have been implemented and where the patterns have been directly imported and adapted using an existing modeling tool. The reuse of model fragments is already supported by the OLED[3] (OntoUML Light Editor), but currently focusing on general patterns and anti-patterns included in its underlying library.

During the application of **Meas** in the PingER domain it was not evident how to explicitly represent the dimensions that qualify a measure, such as the time dimension and geographic location dimension, which could facilitate the visualization of possible analytical perspectives. But, in fact, considering M-OPL and its domain ontology derivations, it does not seem reasonable to contemplate a multidimensional structure, similar to the representation of the Data Cube Vocabulary [CYGANIAK; REYNOLDS; TENNISON, 2014], where facts (measurements) and dimensions (associated concepts) are at the core of the model. Although recognizing the long-term importance of multidimensional models for analytical processing (and, of course, for exploration and aggregation of measurements or statistical data) they serve a different purpose: to make explicit the analytical possibilities associated to the data. But, as such, this type of representation do not constitute a real conceptual model associated to a domain, as it does not usually support the representation of existing relations among concepts, their interdependences and other rules that constitute the rich semantics of the real world conceptualizations.

From the solution found to represent the time and location aspects associated with Measurement in PingER domain, it is possible to conclude that M-OPL can

[3] https://code.google.com/p/ontouml-lightweight-editor/

represent the temporal aspect, treating it as property of the Measurement Relator. But for other characterizations related to Measurement, it seems a better solution to represent them as new concepts, extending some of the patterns proposed in M-OPL.

5. Conclusion

In this paper, we have presented the application of M-OPL to the network performance measurement scenario, in order to derive a new version of PingER ontology [SOUZA *et al.*, 2014] so that we could take advantage of the semantic richness of measurements and their associated concepts when interoperating PingER data with other data as linked open data on the Web.

The benefits of applying M-OPL brought to the development of the new version of PingER ontology were: (i) decreasing the possibility of inconsistencies and ambiguities, since the basic patterns of the M-OPL have been developed following a largely explored theory based on UFO; (ii) acceleration of the ontology development process, as the patterns application process has proven to be effective and easy to use; (iii) as already stated previously, alignment to the core modeling patterns proposed for the measurement area can facilitate the future publication and linkage of the PingER data with related data sources in the Web.

As future work, we are already experimenting with this new version of the ontology, and we expect to evidentiate that the addition of semantic expressiveness brought to the model can lead to more sophisticated and intelligent applications, compensating the inherent increase of complexity of the ontology structure. Also, further discussion on the multidimensional characteristics of measures would be interesting, and what would be the best representation or derivation from a rich conceptualization such as the one already contemplated by M-OPL constructs.

References

Baltrunas, D., Elmokashfi, A., Kvalbein, A. (2014) “Measuring the Reliability of Mobile Broadband Networks”. In: Proceedings of the 2014 Conference on Internet Measurement Conference, p. 45-58, New York, NY, USA.

Barcellos, M. P., Falbo, R. A., Frauches, V. G. V. (2014) “Towards a Measurement Ontology Pattern Language”. In: 1st Joint Workshop ONTO.COM/ODISE on Ontologies in Conceptual Modeling and Informations Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems, Rio de Janeiro, Brazil.

Cottrell, L. (2001) “Internet End-to-end Performance Monitoring”, Available in: http://www-iepm.slac.stanford.edu. Last Access: July 2015.

Cyganiak, R., Reynolds, D., Tennison, J. (2014) “The RDF Data Cube Vocabulary”. In: W3C Recommendation. Available in: http://www.w3.org/TR/vocab-data-cube/. Last Access: July 2015.

Falbo, R. A., Barcellos, M.P., Nardi, J. C., Guizzardi, G. (2013) “Organizing Ontology Design Patterns as Ontology Pattern Languages”. In: 10th European Semantic Web Conference - ESWC 2013 - Semantics and Big Data, Montpellier. The Semantic Web: Semantics and Big Data - LNCS, 2013, v. 7882. p. 61-75, Springer.

Finkelstein, L., Leaning, M. S. (1984) “A Review of the Fundamental Concepts of Measurement”. In: Measurement, v. 2, issue 1, p. 25-34, January-March 1984.

Frauches, V. G. V. (2014) “Uma Abordagem Baseada em Ontologias para Obtenção de Indicadores a partir de Dados Abertos”, MSc. Dissertation, Department of Computer Science, Federal University of Espírito Santo, Vitória, ES, Brazil.

Guizzardi, G. (2005) “Ontological Foundations for Structural Conceptual Models”, PhD Thesis, Centre for Telematics and Information Technology (CTIT), no 05-74, The Netherlands.

Hobbs, J. R., Pan, F. (2006) “Time Ontology in OWL”. In: W3C Working Draft, v. 27, p. 133, September 2006.

Rao, S. (2010) “Monitoring and Measurement in the Next Generation Technologies”. Available in: ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/future-networks/projects-moment-factsheet_en.pdf..

Salvador, A., de Vergara, J. L., Tropea, G., Blefari-Melazzi, N., Ferreiro, Á., Katsu, Á. (2010) “A Semantically Distributed Approach to Map IP Traffic Measurements to a Standardized Ontology”, In: International Journal of Computer Networks & Communications (IJCNC), v. 2, no. 1, p. 13-31.

Scherp, A., Franz, T., Saathoff, C., Staab, S. (2012) “A Core Ontology on Events for Representing Occurrences in the Real World”, In: Multimedia Tools and Applications, v. 58, issue 2, p. 293-331.

Souza, R. F., Cottrell, L., White, B., Campos, M. L., Mattoso, M. (2014) “Linked Open Data Publication Strategies: Application in Networking Performance Measurement Data”, In: The Second ASE International Conference on Big Data Science and Computing, Stanford. ASE BIGDATA / SOCIALCOM / CYBERSECURITY Conference.

Vatant, B., Wick, M. (2012) “Geonames ontology”, Available in: http://www.geonames.org/ontology/documentation.html, July 2015.

The Multiple Applications of a Mature Domain Ontology

Mara Abel[1], Joel Carbonera[1], Sandro Fiorini[1], Luan Garcia[1], Luiz Fernando De Ros[2]

[1]Informatics Institute, [2]Geoscience Institute
Universidade Federal do Rio Grande do Sul (UFRGS)
PO 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{marabel,jlcarbonera,srfiorini,lfgarcia,lfderos}@inf.ufrgs.br

Abstract. *Ontologies have been growing in importance regarding their reusability for distinct applications, since this allows amortizing the significant cost of development of a knowledge base. Large portions of knowledge models now are modelled as ontologies and these portions are shared through several applications. Considering the immature stage of the methodologies of Ontology Engineering and the considerable short space of time for evolving fully operational domain ontology, few reports of real cases of ontology reuse are found in the literature. This article describes a mature domain ontology for Petrographic description and the several knowledge-based applications that it supports. The ontology development started in the 90's and it is still in evolution, both by extending vocabulary as by improving the rigor of the conceptual modelling approaches. We analyze here the impact that each new application has caused over the ontology, requiring improvements and modifications in the original model.*

1. Introduction

Building a fully operational domain ontology is a long time and resource-consuming effort that can keep a team of professionals dedicated for years in refining and improving the knowledge modelled. The team usually demands professionals of the domain along with knowledge engineers and software analysts, whose combined profiles can cover the requirements of expert knowledge, formal correctness, semantic richness and efficiency, required for such knowledge-based applications.

This effort can be rewarded by the several uses that a heavy domain ontology can support if its development has followed methodological approaches that guarantee a high level of generality and modularity of the modelled ontology. Each possibility of reuse brought by the development of a new knowledge-based system in the same domain can amortize the cost of development and maintenance of the domain ontology.

Much has been said about the advantages of building a well-founded domain ontology regarding the potential software applications that can be supported by ontologies. However, ontology engineering is still a recent area of research, and its technological products are just starting to be delivered and evaluated.

Kop in [Kop 2011] discusses the limitations in adopting an existent domain ontology as the basis for a new knowledge-based application. Different views over the domain and ontological choices driven by diverse goals require significant adaptations on the ontology, which are hard to be accomplished by knowledge engineers. The author claims that the reuse can be assured by the involvement of the domain expert

in the ontology adaptation. Confirming the Kop claim, the adaptation of ontology to support new applications was successful applied in the several Geology projects described here, in this article. Still, the motivation for ontology reuse can go beyond the reuse of an available formal piece of knowledge. Shah [Shah *et al.* 2014] has described a framework to help the reuse of a biomedical ontology with the intention of helping the integration of distinct specialties in Medicine thought a common knowledge-based framework of software. Nevertheless, the cost or utility motivation for ontology reuse and the possibilities of reducing the cost of knowledge-based applications by recycling existent ontologies still face the problems of correctness of the ontology modelling [Guarino & Welty 2002], quality of documentation [Simperl *et al.*2011] and the further modifications of a shared ontology that can impact the maintenance of applications [Tsalapati *et al.* 2009].

Our experience shows that, despite of the cost of developing fully operational domain ontologies, the possibilities of reuse of the artifact outspreads the costs and effort of the development.

In order to contribute to the understanding of the potential uses of domain ontologies in knowledge-based applications, this article analyses the actual uses of a mature domain ontology whose development started on 90's and is being continuously enhanced. We described the commercial and non-commercial software applications and how each new application has affected the original definition of concepts and the improvements that were done in order to keep compatibility and modularity among the several supported software families.

2. The PetroGrapher project

The Petrography domain ontology was the main product of the PetroGrapher project developed by the Intelligent Database Group of Federal University of Rio Grande do Sul, Brazil, from 1995 to 2007 [Silva 1997; Abel 2001; Mastella 2004; Victoreti 2007]. The domain ontology was aimed to organize and represent the Geology vocabulary required to support the quality evaluation of clastic and carbonate petroleum reservoirs through petrographic analysis. An intelligent database application – Petroledge® system[1] - was developed to support the petrographer through the task of reservoir description and interpretation. The original ontology published in [Abel 2001] was a partonomy of 21 geological terms (Figure 1), whose attributes and values added another 1500 terms to the initial model. The terms were structured mainly through the *part of* relationship. The more significant hierarchies refer to the mineral constituents: *Detrital* and *Diagenetic composition* classes and subclasses (not detailed in Figure 1). The concept *Diagenetic composition*, its attributes and domain of possible values are detailed in Figure 2. The Figure illustrates the frame-based formalism adopted in the knowledge representation and exemplifies the level of detail in which the ontology was formalized. The knowledge representation formalism was chosen intended to facilitate the mapping of the concept representation to a relational database model, since the database acts as the repository of the domain ontology. The Figure 2, in particular, shows the attributes *Location* and *Paragenetic Relation*, which express the spatial relationships that a diagenetic mineral has with its neighborhood that can be visually recognized by the geologist. The *Diagenetic composition* concept and attributes are essential for the several interpretation tasks described in the Section 3.

[1] Petroledge, Petroquery, Hardledge, RockViewer and Petrographypedia are trademarks of ENDEEPER Company. The suite of ontology-based applications described in this paper can be known in www.endeeper.com/products.

Basically, petrographic evaluation refers to the formal description of visual aspects of a rock sample, as they appear in naked-eye analysis and under an optical microscopic. Starting from the petrographic features that are discerned, the petrographer infers the possible geological interpretation(s) of the rock, which will strongly influence the method of evaluation of the potential of the geological unit as an oil reservoir. The geologist analyses the physicochemical conditions, called *diagenetic environment*, in which the rock was possibly produced, according to the features that would have been imprinted in the rock by the conditions of this environment.

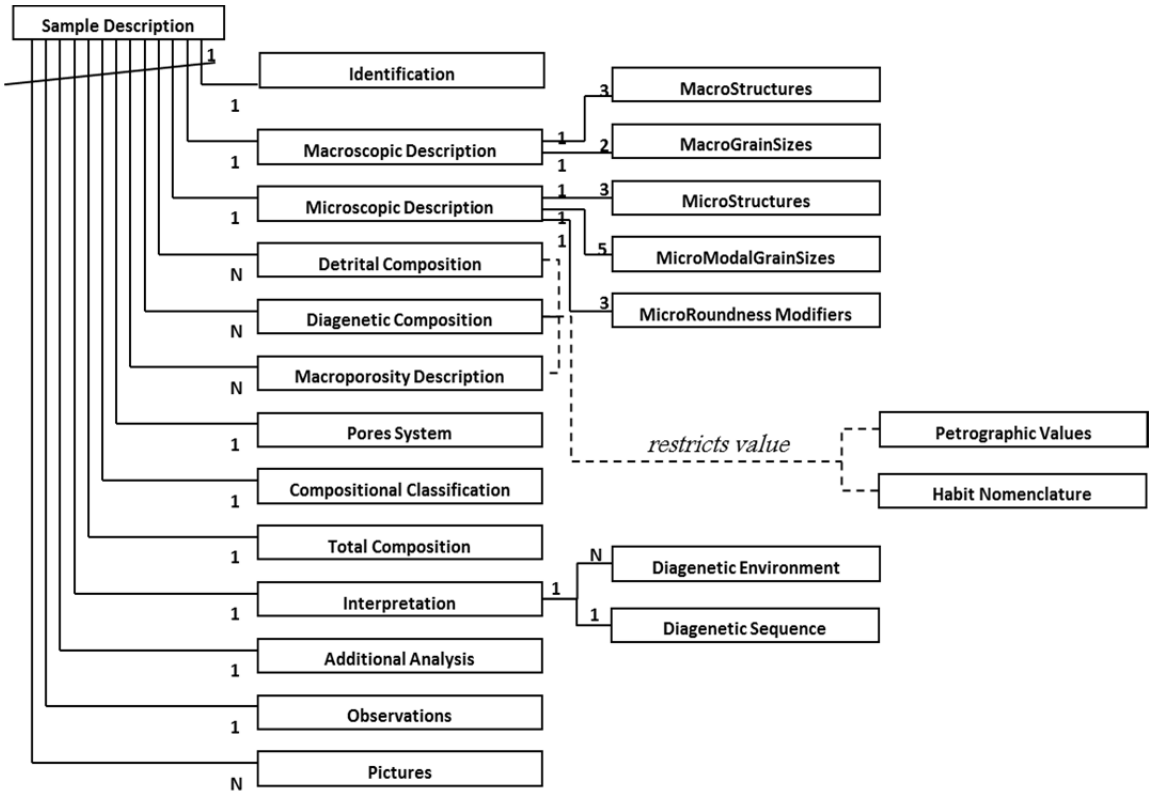

Figure 1. Main concepts of the ontology of Petrography for petroleum reservoir. The boxes describe the concepts and the arcs represent the *part-of* relationship.

The greater challenge in building knowledge application in Geology is that the explicit part of the knowledge that can be expressed through words is just a part of the body of knowledge applied in interpretation. Most of the data relevant for geological interpretation of oil reservoirs consist of visual information that have no formal denomination and are learnt through an implicit process during training and field experience. These features without names constitute the implicit body of knowledge, also called *tacit knowledge* by Nonaka and Takeuchi [Nonaka *et al.* 1995] when referring to the unarticulated knowledge that someone applies in daily tasks but is not able to describe in words. The articulated or *explicit knowledge* that we call ontology refers to the consciously recognized entities and how these entities are organized. Tacit and explicit knowledge should be seen as two separate aspects of knowledge that demands their own representational formalism and not different sorts of it.

Concept Diagenetic-Composition	
Is-a	Object
Part-of	Concept Sample-Description
Mineral Name	one-of **[Diagenetic-Constituent]**
Constituent Set	one-of [Silica, Feldspar, Infiltrated clays, Pseudomatrix clays, Authigenic clays, Zeolites, Carbonates, Sulphates, Sulfides, Iron oxides/hydroxides, Titanium minerals, Other diagenetic constituents]
Habit	one-of **[Habit-Name]**
Amount	range [0.0 - 100.00]
Nominal Amount	one-of [abundant, common, rare, trace]
Location	one-of [intergranular continous pore-lining, intergranular discontinous pore-lining, intergranular pore-filling, intergranular discrete, intergranular displacive, intragranular replacive, intragranular pore-lining, intragranular pore-filling, intragranular discrete crystals, intragranular displacive, moldic pore-lining, moldic pore-filling, oversized pore-lining, oversized pore-filling, grain fracture-filling, grain fracture-lining, rock fracture-filling, rock fracture-lining, concretions/nodules, massive beds/lenses]
Modifier	one-of [dissolved, zoned, fractured, recrystallized]
Paragenetic Relations	one-of [**Covering** <one-of [Diagenetic-Constituent]>, **Covering** <one-of [Detrital-Constituent]>, **Covered by** <one-of [Diagenetic-Constituent]>, **Replacing grain of** <one-of [Detrital-Constituent]>, **Replacing matrix of** <one-of [Detrital-Constituent]>, **Replacing** <one-of [Diagenetic-Constituent]>, **Replaced by** <one-of [Diagenetic-Constituent]>, **Alternated with** <one-of [Diagenetic-Constituent]>, **Engulfing** <one-of [Diagenetic-Constituent]>, **Engulfing** <one-of [Detrital-Constituent]>, **Engulfed by** <one-of [Diagenetic-Constituent]>, **Intergrown with** <one-of [Diagenetic-Constituent]>, **Overgrowing** <one-of [Diagenetic-Constituent]>, **Overgrowing** <one-of [Detrital-Constituent]>, **Overgrown by** <one-of [Diagenetic-Constituent]>, **Expanding** <one-of [Detrital-Constituent]>, **Compacted from** <one-of [Detrital-Constituent]>, **Within intergranular primary porosity, Within intergranular porosity after** <one-of [Diagenetic-Constituent]>, **Within intergranular porosity after detrital matrix, Within intragranular porosity in** <one-of [Detrital-Constituent]>, **Within intracrystalline porosity in** <one-of [Diagenetic-Constituent]>, **Within moldic porosity after** <one-of [Detrital-Constituent]>, **Within moldic porosity after** <one-of [Diagenetic-Constituent]>, **Within shrinkage porosity of** <one-of [Detrital-Constituent]>, **Within shrinkage porosity of** <one-of [Diagenetic-Constituent]>, **Within grain fracture porosity in** <one-of [Detrital-Constituent]>, **Within rock fracture porosity in** <one-of [Detrital-Constituent]>]
Paragenetic Relation Constituent Set	one-of [Silica, Feldspar, Infiltrated clays, Pseudomatrix clays, Authigenic clays, Zeolites, Carbonates, Sulphates, Sulfides, Iron oxides/hydroxides, Titanium minerals, Other diagenetic constituents, Detrital quartz, Detrital feldspar, Plutonic rock fragments, Volcanic rock fragments, Sedimentary rock fragments, Metamorphic rock fragments, Micas/chlorite, Heavy minerals, Intrabasinal grains, Detrital matrix, Other detrital constituents]

Figure 2. A detail of the attributes and domain values of the Diagenetic Composition concept represented in the ontology. The lists [Diagenetic-Constituent] and [Habit- Name] describe the specialized vocabulary that describes mineral names and formats of minerals modelled in a separated way for a question of modularity and reusability.

The Petroledge application was conceived in order to allow a user with a medium level of expertise to describe petrographic features in his/her own level of technical language. The system has the role of applying knowledge to recognize, within those ontologically described features, the items that can serve as diagnostic cues for higher levels of expertise in interpretation, in some imitation of a process of visual interpretation (but with even images being described symbolically). In order to achieve that, the knowledge model represents the connection between the features described using ontological vocabulary and those no-named features utilized by the experts to support interpretation. In other words, the model explicity represents the way in which the expert would *see* the same features seen and described by the user with support of ontology. The knowledge acquisition process and the way in which the knowledge was modeled and implemented in Petroledge system are described in [Abel *et al.* 1998].

3. Ontology-based applications

The long-term effort of building a detailed domain ontology in Petrography had the aim of developing a software application to support the highly specialize task of quality evaluation of petroleum reservoir. Petroledge features include an optimized support for the petrographic description of clastic and carbonate reservoirs and other sedimentary rocks. The system guides sample description, according to a systematic order, allowing the standardization and easy access to petrographic terminology for all aspects of description. The user will produce a structured description of the rock under analysis according to the knowledge model. The knowledge base is composed by the ontology and a set of distinct representational formalisms that describe the scheme of a description and the inferential knowledge applied by problem-solving methods [Gómez-Pérez & Benjamins 1999]. Each description is stored as a set of tuples of concept-attribute-value or any logical combination of concept-attribute-value. Records within a relational database are further processed by several problem-solving methods, each one intended to extract geological interpretation, such as, rock provenance, diagenetic environment of rock formation, original rock composition before diageneses, and others. This simple structure (frames + inferential relationships) is the base for supporting multiple applications.

The more powerful inferential formalism applied by Petroledge is the *knowledge graph*, which plays the role of a rule type (in the sense defined in [Schreiber *et al.* 2000]) in defining the inference paths of the problem-solving process. They were built as an AND/OR tree, where the root represents the interpretation and the leaves are instances of no-named visual features. By its side, each no-named feature is associated to a set of terms of the ontology that better describe the visual aspect of that evidence (Figure 3). This aggregate structure of knowledge and its cognitive significance was firstly defined as a *visual chunk* by [Abel 2001]. The k-graph as a whole represents how much each feature influences the choice of some particular interpretation as a solution for the interpretation problem. It also provides a connection between the expert-level knowledge and the shared ontology applied by the professionals on communication and daily tasks. A weight assigned to each feature assets the relevance of that feature to a particular geological interpretation. Twelve k-graphs represent the knowledge required by Petroledge system to automatically interpret the six possible diagenetic environments for clastic reservoirs. The reasoning mechanism of the Petroledge system exams the description of the user in the database searching for described features that match to each knowledge graph. When the weights of features are enough to support that interpretation, the diagenetic environment and the founded features are shown to the user.

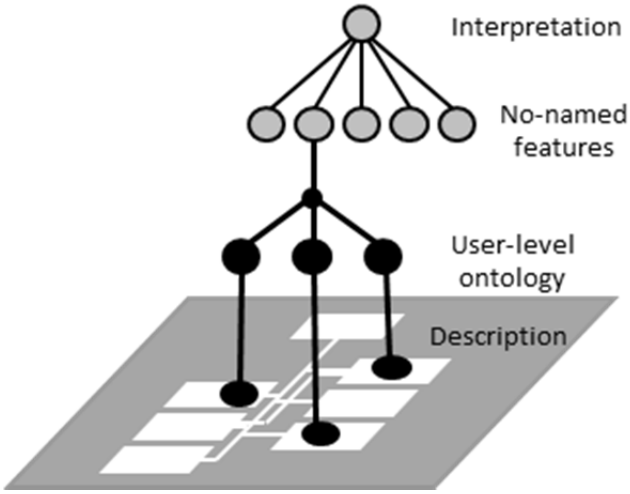

Figure 3. The knowledge graph describes the evidences that support geological interpretation and also links the expert level features to the set of terms in the ontology that describes the content of the evidence.

Figure 4 shows the visual chunk in the petrographic application that describes Diagenetic Dissolution and its internal representation as it is manipulated by the system.

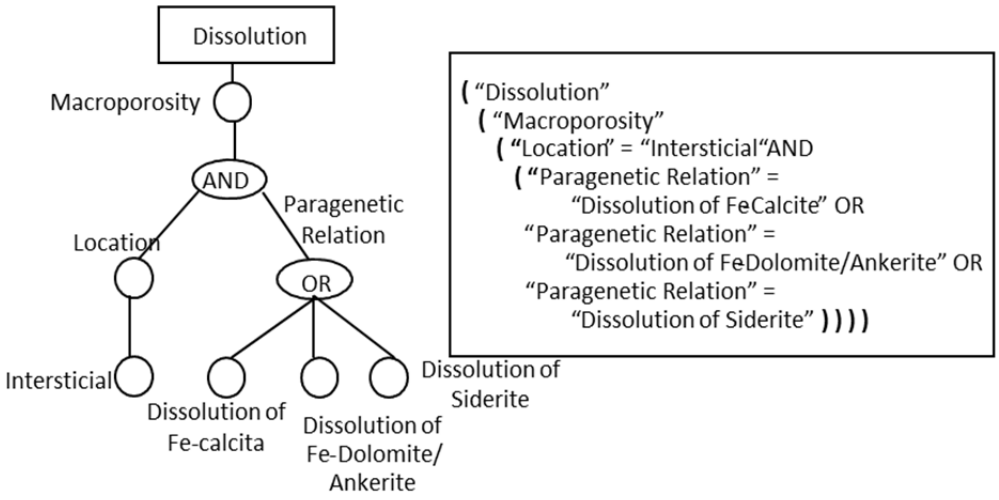

Figure 4. A representation of a visual chunk that describes the inference for Diagenetic Dissolution interpretation and its internal representation as manipulated by Petroledge.

Several other methods of reasoning were developed and applied over the Petrography ontology-based model. Each method requires its own inferential knowledge model and is called or not by the system in an independent way. Compositional classification and provenance interpretation apply numerical methods based on the proportion of minerals. Inferential rules can deconstruct the diagenesis and retrieve the original composition of sediments. Textural classification is based on the proportion of the size of grains. Geological rules can infer the proportion of intrabasinal and extrabasinal sediments.

A further expansion of the ontology model has allowed the modelling of diagenetic sequences, enabling new inference methods to extract the sequence of physicochemical events that has generated a reservoir rock from the spatial relations among mineral constituents [Mastella *et al* 2007]. In order to support that, new concepts describing *events* and *temporal relations* were included in the model and their instances

were defined. In addition, the *paragenetic relations* (showed in Figure 2) that describe mineral constituent associations had their spatial attributes detailed. A set of inference rules describes the relation between the mineral association and the event that has happened with the rock. A reasoning method reads the features described by the user and stored in the database and orders the events that have happened with the rock since the deposition of sediment and later consolidation of the rock. Figure 5 shows the graphical representation of the inference rule that allows ordering the generation of the mineral dolomite as being happened before the generation of mineral anidrite.

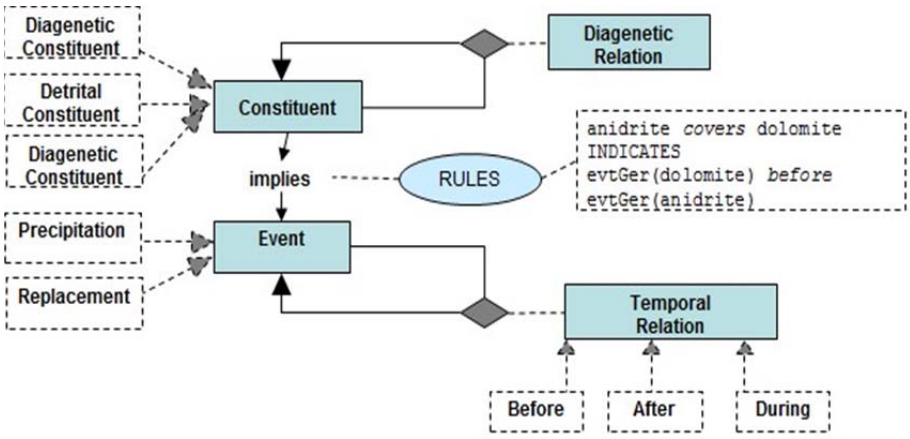

Figure 5. The model of inference rules for extracting sequence of events from the Petrography ontology model.

The flexibility of the ontology model allows each method being based on different inferential knowledge models that are applied by independent modules of software, according to the needs of a particular use of rock data.

Besides the several inference methods that were associated to the Petrography knowledge model, several other applications had been developed getting advantage of the strong and complete formalized vocabulary, even without being part of the Petroledge suite of software.

The Petroquery® application implements a query system over the rock description based on the ontology. Getting advantage of the vocabulary, the application offers to the user his/her own vocabulary for consultation restricting the option of words that are actually present in the database. The user builds SQL consultations by selecting the controlled vocabulary and retrieving the rock descriptions that includes the query arguments. With this support, the geologist can build domain specific consultations like "*Retrieve all rock samples that has dolomite replacing feldspar grains and anidrite within intergranular porosity*".

The controlled vocabulary of the domain ontology was also applied for labeling and indexing microscopic images of rocks in the RockViewer® system, developed in 2010. An editor allows the geologist to associate ontology-controlled text describing images of the rock. After the images being labelled, usually for an experienced petrographer, they are shared through a distributed database to be consulted. The system is used in corporate environment for geologist consultation of the many aspects of rocks that affect the quality of a petroleum reservoir. Figure 6 shows the interaction with RockViewer®. The terms of ontology describing the content of image and used for consultation are highlight in the image label.

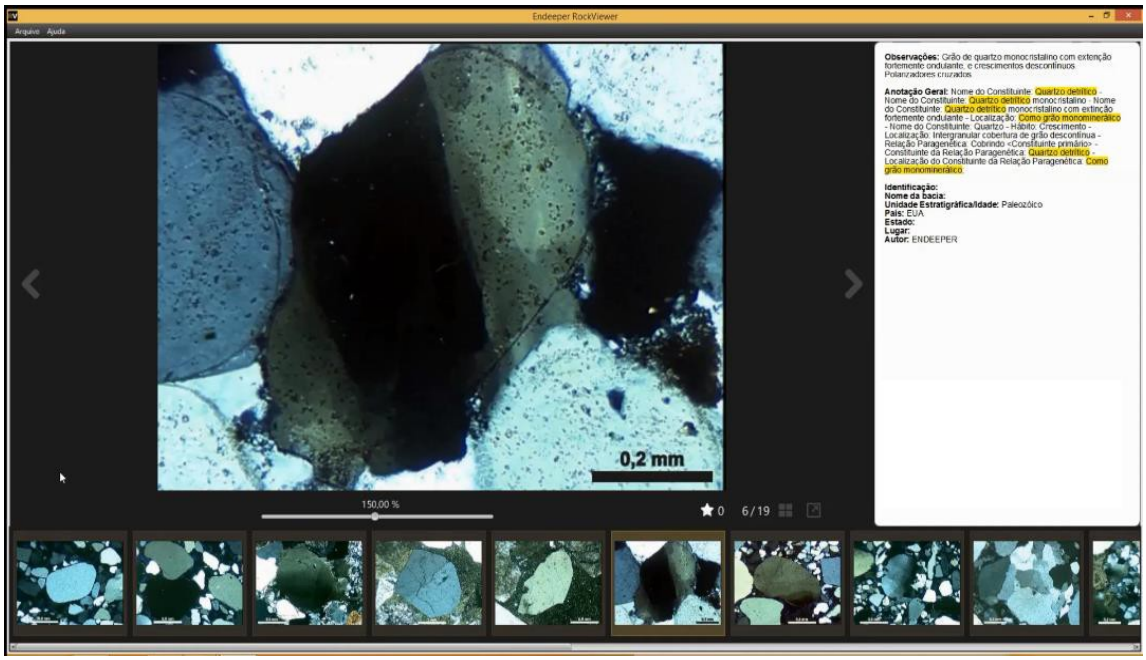

Figure 6. Domain ontology allows to indexing and recovering image content.

The original domain ontology covers the domain of rock-reservoir description. The knowledge schema models the structure of a reservoir description, while the mineral names and characteristics and textural aspects, that constitute the bulk part of ontology, were captured from the more general vocabulary of the Geology community, which supports several other Geology interpretation tasks. Based on this assumption, the ontology of Petroledge was extended to cover all types of rocks and a related knowledge-based application – Hardledge® system - was developed to support mining rock interpretation problems. This 2010's developed ontology was already extended, in 2013, to support the interpretation of magma placement history in sedimentary basins affected by tectonic events.

Other classes of software application can benefit by the reuse of available domain ontology. The web-based application PetrographypediA [Castro 2012] applies the ontology of minerals and their characteristics on microscope to build a visual all-type-of-rock atlas on-line to be freely consulted by the Geology community. As for RockViewer® application, the ontology of Petroledge and Hardledge® was used to label and index rock pictures taken in optical microscope.

A remarkable application of the Petrography domain ontology in the last year is related to the development of conceptual solutions to provide interoperability between reservoir modelling applications along with petroleum chain. The ontology is being used to make explicit the meaning of the geological concepts embedded in the software code and models in order to allow these objects to be recognized and applied to anchor the models of distinct suppliers [Abel et al. 2015b]. This initiative is being conducted by the Energistics[2] consortium in the definition of RESQML interchange standard [King *et al.* 2012]. Also, the PPDM association is applying the well-founded ontology for anchoring the concepts of data models and providing better support for data mapping among different application models [Abel *et al.* 2015a].

[2] ENERGISTICS is a global consortium that facilitates the development, management and adoption of data exchange standards for petroleum industry. RESQML is the data exchange standard for reservoir data. www.energistics.org.

4. The Petroledge Ontology Evolution

The knowledge model of Petrography was initially defined using a frame-based formalism whose general aspect was showed in Figure 2. Two requirements oriented the modelling definition: the understanding of the expert about the information required to produce a qualified rock description and the data management requirements for storage and retrieving a large number of descriptions in a corporate environment. The knowledge acquisition was strongly based on the collection of cases of previous descriptions. As a result, the original model was a flat representation of a *rock description* instead of focusing in the rigid geological concepts and the hierarchy that structure the world in the geologist mind.

The inadequacy of the original model was soon evidenced as much as the reasoning method for diagenetic environment interpretation was developed. To cope with the reasoning, the model was separated in three parts: the knowledge schema of the domain (the partonomy that aggregates each aspect of a rock that needs to be described, showed in Figure 1), the implicit visual knowledge applied by expert in supporting interpretation (later on, it was modelled through visual chunks and knowledge graphs), and the explicitly knowledge or the extensive list of mineral names, textural aspects, lithology nomenclature and the structural relationships that had further grown as the Petrography ontology. Although the knowledge model of rock description and the further extracted visual chunks are still in use in Petroledge and Hardledge® systems, most of maintenance done over the original knowledge model refers to the vocabulary extension and quality improvement of the ontology.

The subsequent evolution was demanded by the interpretation of event sequence that has generated the rock. It was necessary to identify through the domain ontology the upper level classes of the modelled concepts, such as *event*, *temporal relation* and *spatial relation*. This was done by aligning the ontology with other upper ontologies described in literature [Sowa 1995; Scherp *et al.* 2009] and then using the concepts of upper ontology to classify and organize the related concepts in the domain ontology. As a result, the study of the paragenetic relationships described in the Petrography model shows those that represent the spatial relationship between minerals that express the occurrence of an event. Formal definitions of temporal relations based on Allen relations [Allen 1991] were included in the ontology, as well as the definition of events in terms of Geology phenomena. The Allen relations and the definition of diagenetic events allow the extraction and ordering of the events that have transformed the sediments in a consolidated rock from the information described by the user in the rock description.

The RockViewer and PetrographypediA applications were the first Petroledge independent systems that were based on the ontology. As a consequence, these projects have required the ontology rebuilt as an independent artifact, stored in a separated database for further consultation. This reconstruction has produced a new model for the same domain knowledge expressed in the ontology. The rigid concepts (rock and mineral constituent) and their attributes have built the main framework of restructured ontology. New terms were added to expand the domain of application to new kinds of rock and new rock features

The more significant advance for the ontology development came with the use of ontology for improving the interoperability in the petroleum modelling chain by embodying geological explicit concepts and rock properties in RESQML standard. The previously described projects were developed under supervision of the original team of knowledge engineers. For the application into petroleum standards, the ontology needs to be used for several engineers from many distinct software suppliers around the world. The ontology needs to embody all restrictions requested to express the semantic of each

geological term in order to avoid a flexible use with another meaning, which is one of the main sources of errors.

In order to support RESQML integration, each geological concept in the ontology was studied based on the metaproperties proposed by Guarino and colleagues in [Guarino & Welty 2001, 2002; Gangemi *et al.* 2003]. Physical objects, such as *lithological unit*, and amounts of matter, like *rock*, were identified and modeled in a separated way in the geological model. Usually these objects are collapsed or partially merged in the geological models resulting in the main source of problems in reservoir information integration, since many properties related to the substance, such as permeability, are associated to bodies of rock and incorrectly extrapolated by the simulation systems. In addition, the relevant attributes of the concepts that allow defining the identity of each entity were specified as well as their domain of values. The approach of conceptual spaces became the theoretical framework for modeling domain of attributes aiming reusability in other areas of applications into the Geology domain [Fiorini *et al.* 2015]. The ontological analysis of the main concepts of the ontology that are being integrated into RESQML standard can be found in [Abel *et al.* 2015b].

In addition, the problem of scale of analysis that was never an issue for the Petrography domain became central to support applications where the data is generated and consumed in distinct scale of analysis. Basin (105 meters), reservoir (103 meters) and well (10 meters) scales of studies have required that the range of numerical attributes and the symbolic values were extended to cover the new possibilities of the domain.

5. Conclusion

The Petrography ontology has been continuously evolving since it was proposed. From the initially two applications based on a set of twelve concepts, the model embodied a vocabulary as large as 7000 terms split in two idioms which is shared by more than a dozen applications.

This successful grown have been requiring continuous expansion in the number of modelled concepts. Keeping the consistency and integrity of the knowledge base after the inclusion of new concepts have requested periodic restructuring of the ontology organization, sometimes followed by deep changes in the philosophical view that orients the ontological decisions. These changes were especially significant on the first stages of ontology-based application developments and now, when the ontology is going to be integrated to the reservoir interchange standards. The rigor in making explicit the semantic of each vocabulary for a large group of users of diverse specialties driven by many distinct objectives is showing to be a challenge in terms of Ontology Engineering. Some studies about the modularity of ontologies and the possibility of offering specialized partial “views” to users according to their professional profile [Aparicio *et al.* 2014] have indicate some new directions for the ontology evolution.

Acknowledgments: PetrograGrapher project were supported by CNPQ and CAPES. The creation of commercial version of Petroledge and the Endeeper Co. was possible thanks to the grants of FINEP and FAPERGS. We thank Endeeper for providing the software detailed information described in this article.

6. References

Abel M. (2001). "The study of expertise in Sedimentary Petrography and its significance for knowledge engineering (in Portuguese)". In: *Informatics Institute,* UFRGS, Porto Alegre, p. 239.

Abel M., Castilho J.M.V. & Campbell J.A. (1998). "Analysis of expertise for implementing geological expert systems". In: *World Conference in Expert Systems*. Cognizant Communication Offices Mexico City, pp. 170-177.

Abel M., Lorenzatti A., Fiorini S.R. & Carbonera J. (2015a). "Ontological analysis of the lithology data in PPDM well core model". In: *19th International Conference on Petroleum Data Integration and Data Management*. Pennwell Houston.

Abel M., Perrin M. & Carbonera J. (2015b). Ontological analysis for information integration in geomodeling. *Earth Science Informatics*, 8, 21-36.

Allen J.F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, 6, 341-355.

Aparicio J.M.L., Carbonera J.L., Abel M. & Pimenta M.S. (2014). "Ontology View extraction: an approach based on ontological meta-properties". In: *IEEE international conference on tools with artificial intelligence - ICTAI 2014*. IEEE Limassol.

Castro E.S.E.D. (2012). PetrographypediA. The portal of Petrography. URL http://www.Petrographypedia.com/

Fiorini S., Abel M. & Carbonera J. (2015). Representation of part-whole similarity in geology. *Earth Science Informatics*, 8, 77-94.

Gangemi A., Guarino N., Masolo C. & Oltramari A. (2003). Sweetening WordNet with Dolce. *AI Magazine*, 24, 13-24.

Gómez-Pérez A. & Benjamins V.R. (1999). "Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods". In: International Joint Conference on Artificial Intelligence(IJCAI-99), Workshop on Ontologies and Problem-Solving Methods (KRR5) (eds. Benjamins VR, Chandrasekaran B, Gomez-Perez A, Guarino N & Uschold M) Stockolm, Sweden.

Guarino N. & Welty C. (2001). "Identity and Subsumption". In: The Semantics of Relationships. An Interdisciplinary Perspective (eds. Green R, Bean CA & Myaeng SH). Springer Netherlands, pp. 111-126.

Guarino N. & Welty C. (2002). Evaluating ontological Decisions with Ontoclean. C*ommunications of the ACM* 45, 61 – 65.

King M.J., Ballin P.R., Bennis C., Heath D.E., Hiebert A.D., McKenzie W., Rainaud J.-F. & Schey J.C.S.P.E. (2012). Reservoir Modeling: From RESCUE To RESQML. SPE Reservoir Evaluation & Engineering, Society of Petroleum Engineers.

Kop C. (2011). "Domain expert centered ontology reuse for conceptual models". In. Springer Verlag Hersonissos, Crete, Greece, pp. 747-762.

Mastella L. (2004). "An event-based knowledge model for temporal sequence acquisition and representation in Sedimentary Petrography". In: *Informatics Institute,* UFRGS, Porto Alegre.

Mastella L.S., Abel M., Ros L.F.D., Perrin M. & Rainaud J.-F. (2007). "Event Ordering Reasoning Ontology applied to Petrology and Geological Modelling". In: Theoretical /Advances and Applications of Fuzzy Logic and Soft Computing.

(eds. Castillo O, Melin P, Ross OM, Cruz RS, Pedrycz W & Kacprzyk J). Springer-Verlag, pp. 465-475.

Nonaka I., Takeuchi H. & Takeuchi H. (1995). The knowledge-creating company: how Japanese companies create the dynamics of innovation. Oxford University Press, New York.

Scherp A., Franz T., Saathoff C. & Staab S. (2009). "F - A model of events based on the foundational ontology DOLCE+DnS ultralite". In: *K-CAP'09 - 5th International Conference on Knowledge Capture*. ACM SIGART Redondo Beach, CA, United states, pp. 137-144.

Schreiber G., Akkermans H., Anjewierden A., Hoog R.d., Shadbolt N., Velde W.v.d. & Wielinga B. (2000). Knowledge engineering and management:The CommonKADS Methodology. The MIT Press, Cambridge.

Shah T., Rabhi F., Ray P. & Taylor K. (2014). "A guiding framework for ontology reuse in the biomedical domain". In. IEEE Computer Society Waikoloa, HI, United states, pp. 2878-2887.

Silva L.A.L. (1997). "Intelligent database for petroghaphic analysis (In Portuguese)". In: *Informatics Institute*. UFRGS Porto Alegre, p. 114.

Simperl E., Sarasua C., Ungrangsi R. & Burger T. (2011). Ontology metadata for ontology reuse. *International Journal of Metadata, Semantics and Ontologies*, 6,126-145. Inderscience Enterprises Ltd.

Sowa J.F. (1995). Top-level ontological categories. *International Journal of Human Computer Studies*, 43, 669-669. Academic Press Ltd.

Tsalapati E., Stamou G. & Koletsos G. (2009). "A method for approximation to ontology reuse problem". In. Inst. for Syst. and Technol. of Inf. Control and Commun. Funchal, Madeira, Portugal, pp. 416-419.

Victoreti F.I. (2007). "Mapping and documentation of diagnostic visual features for interpretation in a knowledge-based system in Petrography domain". In: *Informatics Institute*, UFRGS, Porto Alegre.

Extended ontologies: a cognitively inspired approach

Joel Luis Carbonera[1], **Mara Abel**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{jlcarbonera,marabel}@inf.ufrgs.br

Abstract. *Within the Knowledge representation community, in general, an ontology is considered as a formal specification of a shared conceptualization. In this sense, ontologies would be constituted of concepts and could be understood as an approach of representing knowledge. In general, ontologies represent concepts in a logical way, adopting the so-called classical theory of representation. Due to this, ontologies can support classification, based on necessary and sufficient conditions, and rule-based reasoning. In this work, we discuss a cognitively inspired approach for extending the knowledge representation capabilities of ontologies. We propose an extended notion of ontologies which incorporates other cognitively plausible representations, such as prototypes and exemplars. The extended ontology has the advantage of supporting similarity-based reasoning, besides the usual logical reasoning.*

1. Introduction

Nowadays, ontologies are widely adopted for *knowledge reusing* and for promoting the *semantic interoperability* among different systems (and humans). Within the knowledge representation community, in general, ontologies are considered as *formal and explicit specifications of a shared conceptualization in a given domain* [Studer et al. 1998]. It is important to notice that, according to this perspective, ontologies would be constituted of *concepts*. In this work, following other works in the field of Artificial Intelligence [Oltramari and Lebiere 2011, Carbonera et al. 2015], we adopt this *conceptualist* [Smith 2004] view about ontologies.

In general, ontologies represent concepts in a *logical* way, assuming the so-called *classical theory* of representation [Murphy 2002], where the concepts are represented by sets of features that are *shared by all the entities* that are abstracted by the concept. Due to this, ontologies are well suited for supporting classification based on *necessary and sufficient conditions* and for supporting *rule-based reasoning*. However, in general, ontologies cannot deal naturally with *typical* features of the concepts [Gärdenfors 2004]; that is, the features that are common to the entities abstracted by the concepts, but that are neither necessary nor sufficient. In this paper, we propose the notion of *extended ontology*, which incorporates other cognitively plausible representations, such as *prototypes* and *exemplars*, and that can support *similarity-based reasoning* (dealing with prototypical effects), besides the usual *rule-based reasoning*.

2. Theories of knowledge representation

Within the Cognitive Sciences there is an ongoing debate concerning how the knowledge is represented in the human mind. According to [Murphy 2002] in this debate there are

three main theories. The *classical theory* assumes that each concept is represented by a *set of features* that are *shared* by *all* the entities that are abstracted by the concept. In this way, this set of features can be viewed as the *necessary and sufficient conditions* for a given entity to be considered an instance of a given concept. Thus, according to this theory, concepts are viewed as *rules* for classifying objects based on features. The *prototype theory*, on the other hand, states that concepts are represented through a *typical instance*, which has the typical features of the instances of the represented concept. Finally, the *exemplar theory* assumes that each concept is represented by a set of *exemplars* of it, which are explicitly represented in the memory. In theories based on prototypes or exemplars, the categorization of a given entity is performed according to its *similarity* with prototypes or exemplars; the instance is categorized by the category that has a prototype (or exemplar) that is more similar to it. There are some works that apply these alternative theories in computer applications [Fiorini et al. 2014].

3. Extended ontologies

As previously discussed, ontologies can be viewed as a paradigm of knowledge representation that adopts the *classical theory* of knowledge representation. In this sense, the classification of instances is performed by checking if they meet the necessary and sufficient conditions of the considered concepts. However, it is well known in the knowledge representation community that, for most of the common sense concepts, finding their necessary and sufficient conditions can be a challenging task [Gärdenfors 2004]. Besides that, according to evidences taken from the research within the Cognitive Sciences [Gärdenfors 2004], for most of the concepts, humans can perform similarity-based classifications, and can consider the typical features of the concepts during the classification process. In this work, we assume that a knowledge representation framework that preserves the flexibility of the human cognition can provide advantages for knowledge-based systems. For example, a system with this capability could classify some individual i as c (where c is some concept) if it is sufficiently similar to a given prototype of c, even when it does not present all the logically necessary features for being considered an instance of c.

In this work, we propose the notion of *extended ontology* ($\chi\mathcal{O}$), which incorporates the conventional features and capabilities of the *classical ontologies* with the possibility of representing typical features of the concepts and of supporting similarity-based reasoning. This proposal adopts some notions originally proposed in our previous works [Carbonera and Abel 2015a, Carbonera and Abel 2015b].

Definition 1. An *extended ontology* ($\chi\mathcal{O}$) is a tuple

$$\chi\mathcal{O} = (\mathcal{C}, \leq, \mathcal{R}, \mathcal{A}, \hookrightarrow, \mathcal{D}, d, \mathcal{I}, v, ext, \mathcal{E}, ex, \mathcal{P}, prot) \quad (1)$$

, where:

- $\mathcal{C}$ is a set $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ of n symbols that represents concepts (or classes), where each c_i is a symbolic representation of a given concept.
- $\leq$ is a *partial order* on $\mathcal{C}$, that is, $\leq$ is a binary relation $\leq \subseteq \mathcal{C} \times \mathcal{C}$, which is reflexive, transitive, and anti-symmetric. Thus, $\leq$ represents a relation of subsumption between two concepts.
- $\mathcal{R}$ is a set $\mathcal{R} = \{r_1, r_2, ..., r_m\}$ of m symbols that represents relations, where each r_i is a symbolic representation of a given relation.

- $\mathcal{A}$ is a set $\mathcal{A} = \{a_1, a_2, ..., a_l\}$ of l symbols that represents properties (or attributes or features), where each a_i is a symbolic representation of a given property.
- $\hookrightarrow$ is a binary relation that relates properties in $\mathcal{A}$ to concepts in $\mathcal{C}$, such that $\hookrightarrow \subseteq \mathcal{A} \times \mathcal{C}$. Thus $a_i \hookrightarrow c_j$ means that the attribute $a_i \in A$ is an attribute of the concept c_j, in the sense that a_i characterizes c_j.
- $\mathcal{D}$ is the set of every possible value of every attribute $a_i \in A$.
- $d\colon \mathcal{A} \rightarrow 2^{\mathcal{D}}$ is a function that maps a given attribute $a_i \in \mathcal{A}$ to a set $\mathcal{D}_{a_i} \subseteq \mathcal{D}$, which is its domain of values. Notice that $D = \bigcup_{i=1}^{l} d(a_l)$.
- $\mathcal{I}$ is a set $\mathcal{I} = \{i_1, i_2, ..., i_p\}$ of p symbols that represents individuals, where each i_j represents a given individual.
- $v\colon \mathcal{I} \times \mathcal{A} \rightarrow \mathcal{D}$ is a function that maps a given individual $i_j \in \mathcal{I}$ and a given attribute $a_i \in \mathcal{A}$ to the specific value $v \in \mathcal{D}$ that the attribute a_i assumes in i_j.
- $ext\colon \mathcal{C} \rightarrow 2^{\mathcal{I}}$ is a function that maps a given concept $c_i \in \mathcal{C}$ to a set $I_{c_i} \subseteq \mathcal{I}$, which is its extension (the set of individuals that it classifies).
- $\mathcal{E}$ is a set $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ of n sets of individuals, where each $e_i \in \mathcal{E}$ represents the set of *exemplars* of a given concept c_i. Notice that $\mathcal{E} \subseteq 2^{\mathcal{I}}$.
- $ex\colon \mathcal{C} \rightarrow \mathcal{E}$ is a function that maps a given concept $c_i \in \mathcal{C}$ to its set of exemplars $e_i \in \mathcal{E}$.
- $\mathcal{P}$ is a set $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ of n prototypes, where each $p_i \in \mathcal{P}$ represents the prototype of a given concept $c_i \in \mathcal{C}$.
- $prot\colon \mathcal{C} \rightarrow \mathcal{P}$ is a function that maps a given concept $c_i \in \mathcal{C}$ to its prototype $p_i \in \mathcal{P}$.

Besides that, for our purposes, the individuals (members of $\mathcal{I}$) are considered as $q - tuples$, representing the respective values of the q attributes that characterize each instance. Thus, each $i_j \in \mathcal{I} = (v(i_j, a_h), v(i_j, a_l), ..., v(i_j, a_p))$, where a_h, a_l and a_p are attributes of i_j.

In our proposal, the sets $\mathcal{E}$ and $\mathcal{P}$ can be *explicitly assigned* to the members of $\mathcal{C}$, or can be automatically determined from the set $\mathcal{I}$. As a basic strategy, a *prototype* $p_i \in \mathcal{P}$ of a given concept $c_i \in \mathcal{C}$, such that $prot(c_i) = p_i$ can be extracted by analyzing the individuals in $ext(c_i)$ and by determining the *typical value* of each attribute of the individuals. If the attribute is numeric, the typical value can be the *average*; if the attribute is *categorical* (or nominal or symbolic), the typical value can be the *most frequent* (the *mode*).

Considering a given $c_i \in \mathcal{C}$, the set of its exemplars, $ex(c_i)$, should be selected in a way that, collectively, its members provide a good sample of the variability of the individuals in $ext(c_i)$. Also, it is important to consider that the exemplars of a concept can be used for supporting the classification of a given individual i and that, for performing this process, it can be necessary to compare i with every exemplar of every concept of the ontology. Thus, it is not desirable to consider all records in $ext(c_i)$ as exemplars for representing c_i, since the computational cost of the classification process is proportional to the number of exemplars that are selected for representing the concepts. Due to this, in our approach we consider that the number of exemplars related to each concept $c_i \in \mathcal{C}$ is defined as a percentage ep (defined by the user) of $|ext(c_i)|$ (where $|S|$ is the cardinality of the set S). This raises the problem of how to select which individuals in $ext(c_i)$ will be consider as the exemplars in $e(c_i)$. We select three main criteria that an individual $i_j \in ext(c_i)$ should meet for being included in $ex(c_i)$: (i) i_j should have a high degree of

dissimilarity with the prototype given by $prot(c_i)$; (ii) i_j should have a high degree of similarity with a big number of observations in $ext(c_i)$; and (iii) i_j should have a high degree of dissimilarity with each exemplar already included in $ex(c_i)$. This set of criteria was developed for ensuring that the set of exemplars in $ex(c_i)$ will cover in a reasonable way the spectrum of variability of the individuals in $ext(c_i)$. That is, our goal is to preserve in $ex(c_i)$ some uncommon individuals, which can be not well represented by $prot(c_i)$, but that represent the variability of the individuals. In our approach, we apply these criteria, by including in $ex(c_i)$ the k first individuals from $ext(c_i)$ that maximize their *exemplariness index*. The exemplariness index is computed using the notion of *density* of a given individual. Regarding some concept $c_i \in \mathcal{C}$, the density of some individual $i_j \in ext(c_i)$, is computed by the function $density\colon \mathcal{I} \times \mathcal{C} \rightarrow \mathbb{R}$, such that,

$$density(i_j, c_i) = -\frac{1}{|ext(c_i)|} \sum_{p=1}^{|ext(c_i)|} d(i_p, i_j) \tag{2}$$

, where d is some dissimilarity (or distance) function (a function that measures the dissimilarity between to entities). Considering this, the set $ex(c_i)$ of some concept c_i, with k exemplars, can be computed by the Algorithm 1.

Algorithm 1: extractExemplars

Input: A concept c and a number h of exemplars
Output: A set $exemplars$ of h instances representing the exemplars of the concept c.
begin
  $exemplars \leftarrow \varnothing$;
  for $j \leftarrow 1$ **to** h **do**
    $eIndex_{max} \leftarrow -\infty$;
    $i_{max} \leftarrow null$;
    foreach $individual \in ext(c)$ **do**
      $density \leftarrow density(individual, c)$;
      $dp \leftarrow d(individual, prot(c))$;
      $med \leftarrow 0$;
      if $exemplars$ *is not empty* **then**
        Compute the distance between $individual$ and each exemplar already included in $exemplars$ and assign to med the distance of the nearest exemplar from $individual$;
      `/* eIndex is the exemplariness index */`
      $eIndex = dp + density + med$;
      if $eIndex > eIndex_{max}$ **then**
        $eIndex_{max} \leftarrow eIndex$;
        $i_{max} \leftarrow individual$;
    $exemplars \leftarrow exemplars \cup \{individual\}$;
  return $exemplars$;

Notice that Algorithm 1 basically selects from $ext(c)$, the individuals that maximize the *exemplariness index*, which is the sum of: (i) distance (or dissimilarity) of the individual from the $prot(c)$; (ii) the *density* of the individual, considering the set $ext(c)$; and the distance (or dissimilarity) of the individual from its nearest exemplar, already included in $exemplars$.

Once a given extended ontology has its concepts, prototypes and exemplars, they can be used by a *hybrid classification engine* for classifying individuals. This component takes as input an individual and provides its corresponding classifications (a set of concepts $classifications \subseteq \mathcal{C}$). Firstly, the classification engine applies a conventional logical reasoning procedure (using the classical part of the extended ontology) for providing a first set of classification hypothesis. Notice that this reasoning process can infer

more than one classification for the same individual. If this process provides, as classifications, concepts that are not specific (if they are not leaves of the taxonomy), the similarity-based reasoning can be used for determining more specific interpretations. The *hybrid classification engine* implements the Algorithm 2.

Algorithm 2: hybridClassification

Input: An individual i.
Output: A set $classification_{set}$ of concepts representing the classifications of i.
begin
  $classification_{set} \leftarrow \varnothing$;
  Perform the logical reasoning for interpreting i, and include the concepts of the resulting classification in $classification_{set}$;
  if *the concepts in* $classification_{set}$ *are not specific* **then**
    $hyp_{set} \leftarrow \varnothing$;
    foreach $c \in classification_{set}$ **do**
      Find the leaves in the taxonomy, whose root is c, and include them in hyp_{set};
    $classification_{set} \leftarrow \varnothing$;
    $MAX \leftarrow -\infty$;
    foreach $c \in hyp_{set}$ **do**
      $app \leftarrow applicability(c, i)$;
      if $app > MAX$ **then**
        $MAX \leftarrow app$;
        $classification_{set} \leftarrow \{c\}$;
      else if $app = MAX$ **then**
        $classification_{set} \leftarrow classification_{set} \cup \{c\}$;
  return $classification_{set}$;

Notice that the Algorithm 2 uses the notion of *applicability*, which, intuitively measures the degree in that a given concept c can be applied as an interpretation for a given observation $individual$. The applicability is computed by the Algorithm 3, using the prototypes and exemplars of the concepts.

Algorithm 3: applicability

Input: A concept c and an instance i.
Output: A value $r \in \mathbb{R}$, which is the degree in that c can be applied as a classification for i.
begin
  $app \leftarrow 0$;
  $pSimilarity \leftarrow sim(i, prot(c))$;
  $eSimilarity \leftarrow 0$;
  Calculate the similarity $sim(i, ex_i)$ between i and each $ex_i \in e(c)$, and assign to $eSimilarity$ the similarity value of the most similar ex_i;
  $app \leftarrow pSimilarity + eSimilarity$;
  return app;

Notice that the Algorithm 3 uses the function sim for measuring the similarity. Intuitively, the similarity is the inverse of the dissimilarity (or distance) between two individuals. Thus, sim has values that are inversely proportional to the values obtained by the function d. Here, we assume that $sim(i_j, i_l) = exp(-d(i_j, i_l))$.

4. Conclusions and future works

In this paper, we propose the notion of *extended ontology*, which integrates the common features and capabilities of conventional ontologies (based on the classical paradigm of knowledge representation) with the capability of dealing with typical features in similarity-based reasoning processes. The extended ontologies can provide more flexibility in classification processes, in the cases that do not have enough information for being classified according to necessary and sufficient conditions.

In future works, we intend to investigate approaches of *instance selection* [Olvera-López et al. 2010] for enhancing our approach for selecting exemplars. Also, we intend to apply the notion of extended ontologies (as well as the algorithms proposed here) for improving the results obtained in [Carbonera et al. 2011, Carbonera et al. 2013, Carbonera et al. 2015] for the task of visual interpretation of depositional processes, in the domain of Sedimentary Stratigraphy. We are also investigating how this approach can be applied for solving other problems, such as ontology alignment. We hypothesize that it is possible to take advantage of the information represented in the form of prototypes and exemplars, as additional sources of evidences in the process of ontology alignment.

References

Carbonera, J. L. and Abel, M. (2015a). A cognition-inspired knowledge representation approach for knowledge-based interpretation systems. In *Proceedings of 17th ICEIS*, pages 644–649.

Carbonera, J. L. and Abel, M. (2015b). A cognitively inspired approach for knowledge representation and reasoning in knowledge-based systems. In *Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 4349–4350.

Carbonera, J. L., Abel, M., and Scherer, C. M. (2015). Visual interpretation of events in petroleum exploration: An approach supported by well-founded ontologies. *Expert Systems with Applications*, 42:2749–2763.

Carbonera, J. L., Abel, M., Scherer, C. M., and Bernardes, A. K. (2013). Visual interpretation of events in petroleum geology. In *Proceedings of ICTAI 2013*.

Carbonera, J. L., Abel, M., Scherer, C. M. S., and Bernardes, A. K. (2011). Reasoning over visual knowledge. In Vieira, R., Guizzardi, G., and Fiorini, S. R., editors, *Proceedings of Joint IV Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies*, volume 776.

Fiorini, S. R., Abel, M., and Carbonera, J. L. (2014). Representation of part-whole similarity in geology. *Earth Science Informatics*, Special Issue on Semantic e-Sciences.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. The MIT Press.

Murphy, G. L. (2002). *The big book of concepts*. MIT press.

Oltramari, A. and Lebiere, C. (2011). Extending cognitive architectures with semantic resources. In *Artificial General Intelligence*, pages 222–231. Springer.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., and Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143.

Smith, B. (2004). Beyond concepts: ontology as reality representation. In *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, pages 73–84.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1-2):161–197.

Unificando a Comparação e Busca de Fenótipos em *Model Organism Databases*

Luana Loubet Borges[1], **André Santanchè**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)

luanaloubet@gmail.com, santanche@ic.unicamp.br

Resumo. *Model Organism Databases (MODs) são largamente utilizados em pesquisas nas áreas médica e biológica. Como cada MOD é usualmente especializado em um tipo de organismo – e.g., peixe-zebra, rato, humano, camundongo – torna-se difícil a busca da mesma característica em organismos distintos para fins de correlação e comparação. Este trabalho apresenta um framework chamado Unified MOD Discovery Engine, cujo objetivo é permitir a correlação e busca de dados de vários MODs, a partir da unificação da sua representação dos dados. Este artigo apresenta o primeiro passo nesta direção, em que foram analisados e comparados os modelos de dados de dois MODs, o ZFIN (peixa-zebra) e MGI (camundongo), como base para a concepção de um modelo unificado. Tal modelo é a base de um grafo interligado, que permitirá ao usuário fazer buscas e comparações de forma unificada.*

1. Introdução e Motivação

Model Organism Databases (MODs) são repositórios específicos para conhecimento biológico [Hedges 2002], cuja definição não é estritamente estabelecida. Consideramos que cada MOD armazena dados sobre um *organismo modelo*, podendo conter seu genótipo e fenótipo, permitindo realizar pesquisas de conhecimento biológico, como genética, desenvolvimento e evolução. Nas últimas décadas o termo "*organismo modelo*" se referia a um pequeno e seleto grupo de espécies, estudadas profundamente em laboratório e ricamente documentadas [Hedges 2002]. Na medida em que os mecanismos para mapeamento genético se tornaram mais acessíveis, o conceito de organismo modelo se expandiu para um conjunto mais amplo de espécies [Hedges 2002].

A comparação de organismos modelo a partir dos seus fenótipos tem um grande potencial na análise e descoberta de correlações entre organismos e fornecerá uma forma eficiente, por exemplo, de identificar genes correlatos candidatos a causar doenças nos diversos modelos [Washington et al. 2009]. Fenótipo é um conjunto de características físicas e comportamentais de um indivíduo, resultante da interação do seu genótipo com o ambiente. Genótipo refere-se à composição genética do indivíduo. Para que esse cruzamento de dados seja possível entre MODs é preciso que eles estejam unificados. No entanto, organismos modelo não são registrados homogeneamente, tendo corriqueiramente, seus dados armazenados em forma de texto livre, além de não ter um modelo unificado, dificultando buscas e comparações automatizadas.

Outro conceito fundamental neste contexto são os *profiles*, que consistem em definir um foco das informações relevantes para realizar buscas, análises e analogia entre

organismos. No contexto de doenças, por exemplo, um *profile* pode ser composto por elementos de descrição do fenótipo da doença e seu genótipo associado. O *profile* torna-se a unidade de busca, isto é, a comparação é feita entre o *profile* buscado – e.g., olho ausente – e aquele recuperado da base de dados. Os fenótipos podem ser associados a ontologias no método Entidade-Qualidade (EQ) [Balhoff et al. 2010], em que a Entidade está contida em uma ontologia específica de organismos, associada a um termo de Qualidade usualmente da ontologia *Phenotype and Trait Ontology* (PATO) [Washington et al. 2009], e.g., *entidade* (olho) e *qualidade* (ausente).

O nosso trabalho visa contribuir neste contexto, através de um framework para unificar MODs heterogêneos e subsidiar a criação de *profiles* que propiciem a comparação de organismos. Ele parte da proposta de um modelo de organismo genérico – criado a partir da análise de modelos para a descrição de fenótipos – que contém dados relevantes para o pesquisador.

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta trabalhos relacionados; a Seção 3 descreve o modelo unificado; a Seção 4 apresenta como será feita a busca; a Seção 5 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

[Washington et al. 2009] utilizaram vários MODs para realizar a integração de genótipos com seus respectivos fenótipos e descobrir genes ortólogos[1] que sofreram mutação em diferentes espécies, resultando em cegueira nos seus portadores. Para este estudo foi preciso gerar um modelo unificado de vários MODs heterogêneos contendo os genes que seriam considerados na comparação, foram escolhidos 11 genes humanos que possuem genes ortólogos em camundongos, peixe-zebra e drosófila, contidos no *Online Mendelian Inheritance in Man* (OMIM), além de genes de camundongos, peixes-zebra e drosófilas obtidos de bases diferentes.

[Washington et al. 2009] obtiveram os seguintes resultados: (i) alelos variantes contém fenótipos mais similares que os demais alelos do mesmo gene; (ii) é possível recuperar genes mutantes responsáveis por fenótipos anômalos a partir da análise de similaridade destes fenótipos; (iii) identificação de genes ortólogos pelo cruzamento de dados de fenótipos em diferentes espécies. Estes resultados não seriam obtidos se fosse feita a comparação apenas com o genótipo, pois esta abordagem apresenta dois problemas principais: (1) as bases genéticas de grande parte das doenças normalmente são desconhecidas; (2) ainda que a base genética seja conhecida, algoritmos de comparação de genes e/ou genótipos são feitos através do alinhamento de sequências; no caso de doenças ocorre uma mutação no gene causador da mesma, tornando tais algoritmos inadequados, pois essa comparação trata genes a partir da similaridade entre as cadeias. Por esta razão, a comparação é feita através dos fenótipos das doenças, neste caso, os sintomas da doença.

[Washington et al. 2009] enfrentaram duas grandes dificuldades: (1) tiveram que criar manualmente um modelo homogêneo de vários MODs utilizados apenas para o *profile* analisado; (2) criaram um *profile* a partir de várias ontologias, selecionando os termos relevantes para a pesquisa. Da mesma forma, vários pesquisadores enfrentam as mesmas dificuldades, tendo que integrar MODs e definir *profiles* manualmente, pois não existe

[1] genes derivados de um ancestral comum que possuem a mesma função em espécies diferentes

ferramenta computacional que construa um modelo unificado a partir de vários MODs distintos e que suporte profiles associados a ontologias.

Phenomicdb (`http://phenomicdb.info/`) é uma ferramenta que realiza a integração de vários MODs para pesquisas com fenótipos [Kahraman et al. 2005]. Comparado com a nossa proposta, a busca realizada é limitada a apenas uma descrição de um item de fenótipo. O diferencial do nosso trabalho é que ele suportará buscas por *profiles* com vários itens descritivos, utilizando diferentes formatos para a representação de fenótipos.

3. Modelo Unificado

Com o objetivo de sanar a dificuldade relatada na seção anterior, este trabalho propõe um framework para realizar a busca e comparação de *profiles* definidos pelo usuário em um conjunto de MODs de forma transparente. O ponto de partida foi analisar dois MODs de referência amplamente usados e citados em trabalhos relacionados – o ZFIN e o MGI – como bases para a proposta de um modelo unificado.

ZFIN é um MOD que contém tanto dados de genótipos quanto fenótipos do peixe-zebra, em que os fenótipos são descritos pelo método EQ citado anteriormente [Sprague et al. 2006, Washington et al. 2009]. O modelo parcial do banco de dados referente a fenótipos do ZFIN é apresentado na Figura 1(a). Uma descrição de fenótipo é formada por um conjunto de declarações (`Phenotype_statement`) envolvendo uma Entidade (`ZFA_term`) e uma Qualidade (`PATO_term`) ligadas a ontologias externas: ZFA (Zebrafish Anatomy Ontology), GO (*Gene Ontology*) e PATO. Entidades e qualidades são generalizadas como termos (`term`) que têm um auto-relacionamento com tipo (e.g., is-part-of), pois pode-se construir uma taxonomia de termos.

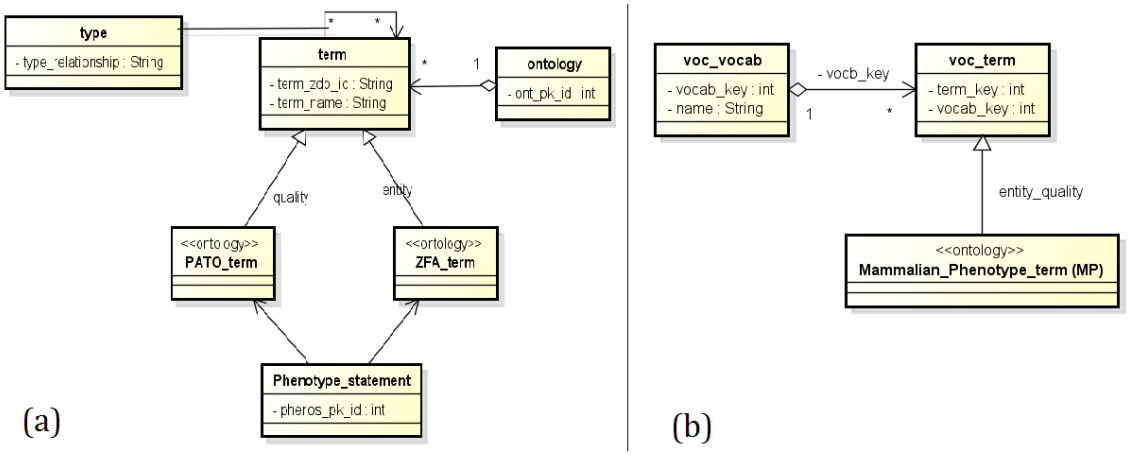

Figura 1. Modelo do banco de dados do ZFIN e do MGI.

MGI é um MOD com dados de genótipos e fenótipos de camundongos [Blake et al. 2003]. A Figura 1(b) retrata um modelo parcial do banco de dados de fenótipos do MGI. A descrição do fenótipo, assim como no ZFIN, é tratada como um conjunto de declarações. Cada declaração corresponde no MGI a um termo (`voc_term`). Cada termo é associado à ontologia *Mammalian Phenotype* que é uma variante da abordagem EQ, pois cada conceito da ontologia já é a composição da Entidade mais a Qualidade

[Smith et al. 2004]. A classe `voc_vocab` correspondente à classe `ontology` do modelo do ZFIN e possibilita o uso de termos de várias ontologias.

A Figura 2 apresenta o nosso modelo unificado, em que um fenótipo (`Phenotype`) é composto por um conjunto de declarações (`Statement`) que correspondem à composição de Entidades e Qualidades, como acontece no `voc_term` do MGI. A classe `Statement_EQ` especializa o `Statement` e é capaz de representar a entidade e a qualidade de forma discriminada como faz o ZFIN (classe `term`). A classe `voc_vocab` do MGI e `ontology` do ZFIN correspondem à classe `Ontology` no modelo proposto. Além disso, as classes `Statement`, `Entity` e `Quality` possuem um auto-relacionamento para registrar sinônimos. A classe `Profile` é formada por um `Phenotype`. Futuramente o `Profile` será integrado com informações de genótipos também.

Os modelos apresentados do ZFIN e do MGI refletem o banco de dados relacional original de ambos. Entretanto, nosso modelo unificado é baseado em uma estrutura de grafos e por isso mapearemos os modelos para um banco de dados de grafos de propriedades [Robinson et al. 2013] fazendo com que cada classe vire um nó, os relacionamentos serão arestas e os atributos das classes viram propriedades dos nós e/ou arestas. O mesmo acontece com o modelo proposto neste trabalho.

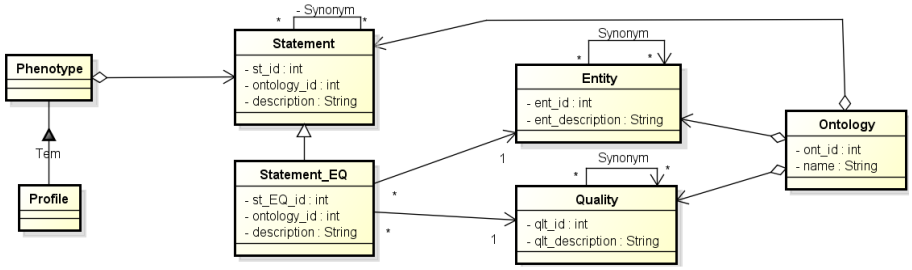

Figura 2. Modelo proposto para a ferramenta *Unified MOD Discovery Engine*.

4. Busca baseada em Profile

Esta seção descreve a arquitetura que projetamos para a realização de uma busca unificando diferentes MODs, em que há um esforço extra para tratar a representação heterogênea dos dados de cada base, já que eles não são homogêneos. Descrições de fenótipos podem ser encontradas em formatos distintos, como textos livres (o que dificulta o uso computacional), C/CS (que é uma forma de descrição semi-estruturada), Entidade-Qualidade (EQ) e uma variante dele que chamaremos de *EQ composto* (tal como no MGI). Como exemplo das formas de descrições, temos que no OMIM as descrições são em texto livre, no MGI são em EQ *composto* e no ZFIN são em EQ.

O nosso sistema propõe a unificação da busca e comparação em MODs distintos. A busca/comparação é feita a partir de uma interface unificada, que fornecerá uma visão homogênea das informações, independentemente de como elas estão armazenadas nos seus MODs de origem.

Tomando o caso descrito por [Washington et al. 2009] como base de pesquisa em vários MODs, apresentaremos a nossa arquitetura através de um exemplo de uma consulta feita no ZFIN e MGI. Ao fazer uma busca no ZFIN pelo fenótipo *lens decreased size* são

retornados vários genes associados a esse fenótipo, entre eles, o gene Pax6b. Esse fenótipo é descrito por meio de sua entidade (*lens*) separada de sua qualidade (*decreased size*).

Ao realizar a mesma busca pelo fenótipo *lens decreased size* no MGI são retornados vários genes, entre eles o gene pax6 que causa microftalmia, que refere-se ao olho pequeno. Mas a interpretação não é tão trivial pois o sistema não retorna o fenótipo exatamente como ele foi buscado. O fenótipo microftalmia tem o sinônimo *lens decreased size* que foi buscado anteriormente. Essas descrições de fenótipos no MGI estão em EQ *composto*.

Ao interligar essas informações do ZFIN e MGI obtemos os genes que causam doenças que levam a cegueira no zebrafish e no camundongo. Essas informações são úteis para realizar pesquisas sobre essa doença também em humanos, já que o gene causador da cegueira em humanos é o PAX6 ortólogo aos genes do peixe-zebra e camundongo.

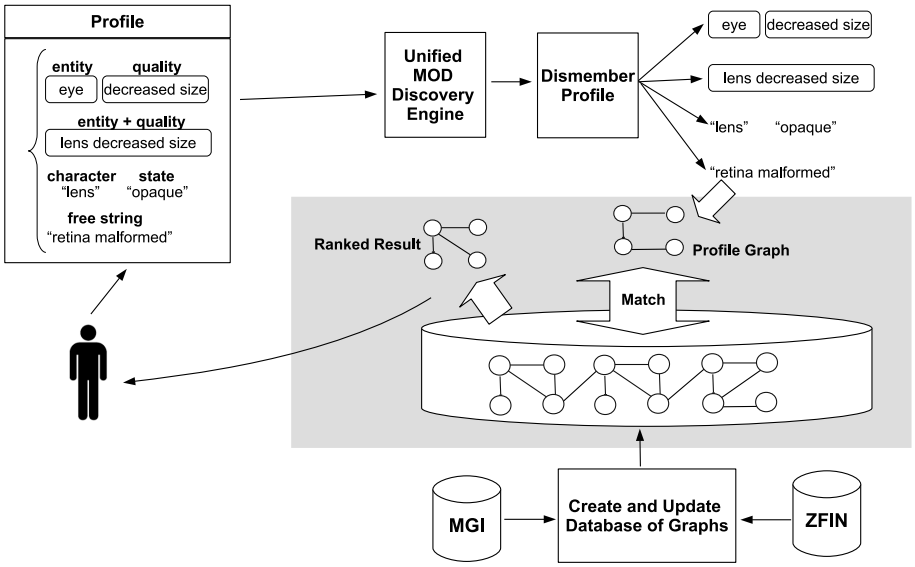

Figura 3. Arquitetura da nossa proposta.

A Figura 3 representa a nossa proposta. O usuário interagirá com a ferramenta na criação do *profile* que é dado como entrada. Neste caso, cada linha corresponde a uma descrição de fenótipo dada pelo usuário, podendo ser em texto livre, EQ, entre outras. Em seguida, a nossa ferramenta terá acesso a um banco de dados de grafos criado previamente que importa as informações contidas no ZFIN e MGI referentes a fenótipo. O nosso framework *Discovery Engine* executará algoritmos de *match* para comparar e analisar profiles. Para tornar possível essa comparação é necessário desmembrar o *profile* em unidades básicas que descrevem o fenótipo (*dismember profile* na Figura 3). Sobre estes itens serão aplicados algoritmos para análise de similaridade para busca e comparação de *profiles*. Como resultado da busca, a ferramenta gera um grafo contendo resultados com informações do ZFIN e MGI ranqueadas por similaridade. O *Profile Graph* da Figura 3 corresponde à representação do profile na forma de grafo, a ser confrontado com as descrições de fenótipos em banco de dados de grafos. Além de importar dados do ZFIN e MGI o banco de dados de grafos também será usado para interligá-las e melhorar o resultado das comparações.

Para realizar a busca no banco de dados através do *profile* utilizaremos métricas

de similaridade também usadas por [Washington et al. 2009]: *Information Content* (IC), métricas semânticas de similaridade e análise de sobreposição [Mistry and Pavlidis 2008].

5. Conclusões

Pesquisadores precisam cruzar dados de vários organismos e recorrem a diversos MODs, contendo diferentes representações de dados, dificultando a interligação dos mesmos. Neste trabalho nós apresentamos um modelo unificado para representação de fenótipos – baseado na análise de dois MODs, o ZFIN e o MGI – bem como o projeto do framework *Unified MOD Discovery Engine*, que permitirá ao usuário realizar buscas por descrições de profiles de organismos em MODs distintos de forma unificada.

Como trabalhos futuros pretendemos implementar o *engine* cujo projeto foi apresentado neste artigo e estender a proposta para outros MODs, como OMIM (humanos), RGD (ratos), Flybase (moscas), entre outros. Além de integrar informações de genótipos que ainda não estão sendo consideradas.

Agradecimentos. Este trabalho foi parcialmente financiado pela AGENCIA, FAPESP/Cepid em Engenharia e Ciência da Computação (2013/08293-7), o Instituto Microsoft Research FAPESP Virtual (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), , INCT em Web Science e subvenções individuais do CNPq.

Referências

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., and Vision, T. J. (2010). Phenex: ontological annotation of phenotypic diversity. *PLoS One*, 5(5):e10500.

Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., Eppig, J. T., Group, M. G. D., et al. (2003). Mgd: the mouse genome database. *Nucleic acids research*, 31(1):193–195.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849.

Kahraman, A., Avramov, A., Nashev, L. G., Popov, D., Ternes, R., Pohlenz, H.-D., and Weiss, B. (2005). Phenomicdb: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, 21(3):418–420.

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, 9(1):327.

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph databases*. O'Reilly.

Smith, C. L., Goldsmith, C.-A. W., and Eppig, J. T. (2004). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7.

Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S., et al. (2006). The zebrafish information network: the zebrafish model organism database. *Nucleic acids research*, 34(suppl 1):D581–D585.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11):e1000247.

An Application Ontology to Support the Access to Data of Medical Doctors and Health Facilities in Brazilian Municipalities

Aline da Cruz R. Souza, Adriana P. de Medeiros, Carlos Bazilio Martins

Department of Computing - Science and Technology Institute, Fluminense Federal University (UFF), Rio das Ostras – RJ – Brazil

acrsouza@id.uff.br, adriana@vm.uff.br, carlosbazilio@id.uff.br

Abstract. *The Web of Data is a global data space based on open standards. However, it is still far from reality found in websites: unstructured and disconnected data focused on human understanding. This work aims to mitigate this problem for a portion of data in the health area, data about medical doctors and health facilities. Thus, this paper presents an application ontology designed to accurately represent such data and some examples of instances and queries, which can be used on the development of applications in order to provide precise information for Brazilian citizens.*

1. Introduction

According to the Google Trends tool, the search volume for the keyword "medical doctor" from the Google search engine had index 60, on a 0 to 100 scale, in the period from 2008 to 2013. The index shows the number of times a keyword was searched on Google in relation to the total number of searches performed in the period. It suggests a considerable use of services like Google to search for data about medical doctors and related terms in the health area. Commonly, Brazilian websites returned from such queries contains unstructured and often incomplete data, mixed with other types of content, such as advertising, hindering the efficient use of these data by citizens. Moreover, published data are, usually, for human processing, which makes hard the reuse of such data in applications.

The Semantic Web (SW) is a Web of Data - dates, names, and any other data that could be conceived. Its technologies (RDF, OWL, SPARQL, etc.) provide an environment in which an application can query this data, make inferences using vocabularies, etc. The set of interrelated data on the Web of Data is called Linked Data (LD) [World Wide Web Consortium (W3C), s.d.]. It allows interconnections to be defined between items in different data sources, aiming a unique global information space [Heath & Bizer, 2011]. The LD principles introduced by Tim Berners Lee [Bizer, et al., 2009] are the following: use URIs as names for things; use HTTP URIs, so that people can look up those names; provide useful information, using standards (RDF, SPARQL); and include links to other URIs for discover more things.

This work presents an application ontology designed to accurately represent data of medical doctors and health facilities of Brazilian municipalities. Instances of the ontology classes were created following LD principles and will become available through a public repository. Some examples of queries that a SW application could perform to

aid citizens to access these semantically structured data are presented. Then, final considerations are provided about the use of the SW technologies in the development of applications to allow access to such data.

2. The Proposed Ontology

Application ontologies describe concepts of a domain and specific tasks for implementing systems, the practical part [Guarino, 1997]. The proposed ontology was created following the "Ontology Development 101" [Noy & McGuiness, 2001] guide. Domain was defined as medical doctors, requiring data about medical doctors and health care facilities of Brazilian municipalities. The scope of the ontology was determined by drafting the following list of Competency Questions (CQ) for which the repository should provide answers: CQ1: *What is the specialty of a particular doctor?*; CQ2: *What are the workplaces of a doctor?*; CQ3: *Does a doctor have more than one CRM?;* CQ4: *Do medical doctors have more than one specialty?*; CQ5: *What are the characteristics considered by a citizen when choosing a doctor?*; CQ6: *Is there any hospital in my neighborhood with a particular specialty?;* and *CQ7: What are the available medical specialties in a given clinic?*.

Thereafter, searches were performed in the DAML (http://www.daml.org/) and Schemapedia (http://schemapedia.com/) repositories, in order to locate validated ontologies that could be integrated to this work by reusing their terms. None of them completely met the work needs, possibly because it is a very specific theme. In this ontology version some terms and resources provided by FOAF (http://xmlns.com/foaf/spec), Geonames (http://www.geonames.org/) and DBpedia (http://dbpedia.org/) regarding to cities were used. Thus, the ontology contains internal terms, data, and references to resources from other repositories.

2.1. Classes and Properties

The ontology was constructed using the OWL 2 DL language, with 98 classes and 656 axioms identified by the prefix *med*. Figure 1 depicts the main classes of the ontology. Vertices are classes and edges are relationships between classes. Dashed arrows represent object properties while continuous arrows represent subclasses.

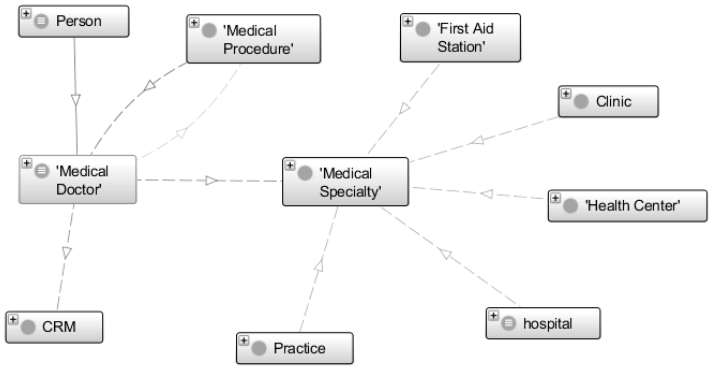

Figure 1. Graph visualization of the ontology.

The class *med:MedicalDoctor* describes a medical doctor, defined as a subclass of *foaf:Person* and as an equivalent class to *dbo:Medician* class. The class *med:CRM*

represents a CRM (registry at the Regional Council of Medicine), defined as a subclass of the class *foaf:Document*. *med:MedicalSpecialty* represents a medical specialty and has medical specialties as subclasses, such as *med:Dermatology*. *med:SurgicalSpecialty* describes a surgical specialty, defined as subclass of *med:MedicalSpecialty*, and has surgical specialties as subclasses, such as *med:SpineSurgery*. The subclasses of *med:MedicalSpecialty* and *med:SurgicalSpecialty* allow answers for the CQ1. The classes *med:Clinic*, *med:Practice*, *dbo:Hospital*, *med:HealthCenter* and *med:FirstAidStation* describe workplaces of a doctor, answering the CQ2. Finally, the class *med:MedicalProcedure* represents a medical procedure that a doctor can perform.

The datatype property *foaf:name* describes the name of something, *dbo:address* represents the address of a place, *dbo:date* corresponds to the date of some event, *dbo:status* is used to represent the status of a CRM (active, inactive, etc.), *med:UF* represents the Unity of Federation in which it was issued, and *dbo:number* describes the number. The object property *med:HasCRM* relates a medical doctor to a CRM and is characterized as inverse functional, which guarantees that an instance of the class CRM relates with a single instance of the class *med:MedicalDoctor*. Regarding the CQ3, the answer is "Yes, it is possible.", but a CRM must be associated with only one doctor. The property *med:WorkplaceOf* relates a person (*foaf:Person*) to his/her workplace (*dbo:Place*) and has an inverse property, the *med:WorksAt*, where the domain and range are reversed. Likewise, *med:PerformedBy* also has an inverse property called *med:Perfoms*. This property represents a medical procedure (*med*:*MedicalProcedure*) performed by a medical doctor (*med:MedicalDoctor*). *The property med:HasExpertise* indicates that a doctor has an expertise (*med:MedicalSpecialty*) and *med:HasSpecialty* relates a health unit to a medical specialty.

Restrictions work as basis for the inferences made by the reasoner, defining which features an instance must have to belong to a certain class. For example, an individual is associated to *med:MedicalDoctor* class when he/she has at least one CRM and has expertise on at least one medical specialty. Regarding CQ4, the answer is "Yes, it's possible", so no restrictions were made to constrain the number of specialties associated to a doctor. Another example of restriction specified for the classes *med:Clinic*, *med:Practice*, *dbo:Hospital*, *med:HealthCenter* and *med:FirstAidStation* is that these health facilities shall have at least one medical specialty.

2.2. Instances

The instances were created according to information extracted manually from websites of Brazilian private health plans and data sources of the Brazilian government. The main data sources used were: Unimed Medical Guide (http://www.unimed.coop.br/) - where were extracted the medical doctors names, number of CRM and specializations; Consulta CRM (http://www.consultacrm.com.br/) - where can be collected the remaining CRM data through an API; DATASUS (http://cnes.datasus.gov.br/) - where were extracted URIs of pages that describe each health facility; Website of Rio das Ostras prefecture (http://www.riodasostras.rj.gov.br) - where were collected the medical specialties provided by the health facilities; Geonames and DBPedia - where were collected URIs of resources that represent the cities of Rio das Ostras and Macaé focus of this work. For privacy reasons fictitious data were used in this section examples. Listing 1 shows an instance of *med:MedicalDoctor*, specifically the doctor "Sara de Sa".

The object properties *#HasCRM* (*med:HasCRM*) and *#HasExpertise* (*med:HasExpertise*) were defined as required conditions for association with the *med:MedicalDoctor*. The property *#WorksAt* (*med:WorksAt*) relates this instance to instances of health facilities in which the medical doctor provides services.

```
<ClassAssertion>
    <Class IRI="#MedicalDoctor"/>
    <NamedIndividual IRI="#Sara_de_Sa"/>
</ClassAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#HasCRM"/>
    <NamedIndividual IRI="#Sara_de_Sa"/>
    <NamedIndividual IRI="#CRM_841631"/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#HasExpertise"/>
    <NamedIndividual IRI="#Sara_de_Sa"/>
    <NamedIndividual IRI="#REC_Pediatria"/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#WorksAt"/>
    <NamedIndividual IRI="#Sara_de_Sa"/>
    <NamedIndividual IRI="#PS_Familia_Rocha"/>
</ObjectPropertyAssertion>
```

Listing 1. Example of instance of *med:MedicalDoctor*.

Listing 2 shows an instance of *med:HealthCenter* representing the Família Rocha Health Center. Note that the property *med:WorkplaceOf*) is inferred by the reasoner from its inverse property *med:WorksAt*. Also worth highlighting relationships with resources located in the external repository Geonames via *foaf:based_near*.

```
<ClassAssertion>
    <Class IRI="#HealthCenter"/>
    <NamedIndividual IRI="#PS_Familia_Rocha"/>
</ClassAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#HasSpecialty"/>
    <NamedIndividual IRI="#PS_Familia_Rocha"/>
    <NamedIndividual IRI="#REC_Pediatria"/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#WorkplaceOf"/>
    <NamedIndividual IRI="#PS_Familia_Rocha"/>
    <NamedIndividual IRI="#Sara_de_Sa"/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty abbreviatedIRI="foaf:based_near"/>
    <NamedIndividual IRI="#PS_Familia_Rocha"/>
    <NamedIndividual IRI="http://sws.geonames.org/3451205/"/>
</ObjectPropertyAssertion>
```

Listing 2. Example of instance of *med:HealthCenter*.

3. Examples of Queries

The following examples show some queries that a SW application could perform from the repository to present useful information for its users. For instance, consider that a woman wants to search for a female gynecologist. Queries like these help to answer the CQ5. The query in SPARQL and its result are shown in Figure 2. This query searches the name, the address and the phone of the medical doctor's workplace. It searches an

individual of the type “medical doctor”, whose sex is female, i.e., whose sex is related to the string “FEMININO” by *dbo:sex* and whose medical specialization is related, by the property *med:HasExpertise,* to the instance *med:REC_Ginecologia_e_Obstetricia*. The workplace, represented by *?health_unit*, is associated to name, address and phone by the properties *foaf:name*, *dbo:address* and *foaf:phone*, respectively. The FILTER clause attends an application specification of returning only health facilities of the type *med:Clinic* or *med:HealthCenter*.

```
SELECT  ?name_doctor ?health_unit_name ?health_unit_address ?health_unit_phone
   WHERE
   {
        ?medico         foaf:name ?name_doctor; a med:MedicalDoctor ;
                        med:WorksAt ?health_unit ;
                        med:HasExpertise med:REC_Ginecologia_e_Obstetricia ;
                        dbo:sex "FEMININO"^^xsd:string.
        ?health_unit    a ?tipo ; foaf:name ?health_unit_name ;
                        dbo:address ?health_unit_address ;
                        foaf:phone ?health_unit_phone .
        FILTER(?tipo IN (med:HealthCenter, med:Clinic))
   }
```

name_doctor	health_unit_name
"KARIN DA PENHA BARROS"	"POSTO DE SAUDE DA CIDADE"
health_unit_address	**health_unit_phone**
"RUA SANTA CATARINA, S/N	tel:+55-22-2768-2008

Figure 2. Query workplace information of a female gynecologist.

Suppose now a search for hospitals that perform general surgery near Rio das Ostras city. Figure 3 shows this query that could answer the CQ6.

```
SELECT ?nome_unidade ?endereco_unid_saude ?tel_unid_saude
   WHERE
   {
        ?unidade        foaf:name ?nome_unidade;    a ?type ;
                        med:HasSpecialty med:REC_Cirurgia_Geral ;
                        dbo:address ?endereco_unid_saude ; foaf:phone ?tel_unid_saude ;
                        foaf:based_near <http://dbpedia.org/resource/Rio_das_Ostras> .
                        FILTER(?type IN (dbo:Hospital, med:FirstAidStation))
   }
```

nome_unidade	
"HOSPITAL MUNICIPAL "@pt	
endereco_unid_saude	**tel_unid_saude**
"RUA EUDON MUSTOSA - S/N - PARQUE"	tel:+55-22-2779-6329

Figure 3. Query hospitals near Rio das Ostras, which perform General Surgery.

Such query returns an individual related to the individual *med:REC_Cirurgia_Geral* by the property *med:HasSpecialty*. It shows a link between a resources of the DBpedia and of the local repository by the property *foaf:based_near*. According to the LD recommendations, links with other repositories allow applications to obtain useful information following these links [Heath & Bizer, 2011]. For instance, from this link with DBpedia it is possible obtain other information, e.g., a place description. An application could get this data via HTTP requests sent to a SPARQL endpoint [Sequeda, 2012]. This is a fundamental difference between SPARQL and other query languages such as SQL, which assume that all data being queried are local and conform to a single model. To answer the CQ7, a query returning the medical specialties associated with the clinic would be enough.

4. Conclusions and Future Works

This paper presented an application ontology for describing data of medical doctors and health facilities in a semantic way, in order to facilitate the development of applications for providing access to these data by Brazilian citizens. The data were represented as axioms structured in RDF and expressed from links with Geonames and DBpedia. Finally, some examples of queries that an application could perform were presented.

Many ontologies and vocabularies are available for the health area, such as OMRSE [Brochhausen, et al., s.d.] and those stored in the OBO-Foundry repository [Ashburner, et al., s.d.]. However, most of them describe information that is not provided in this work, like diseases and human body anatomy. The ontology SNOMED CT [IHTSDO, s.d] and the upper ontology UMBEL [Bergman, M. K. & Giasson, F., s.d.] specify terms relating to medical specialties and health facilities. Specifying the relationship between the terms of the proposed ontology and the terms of these ontologies is an ongoing work. Future works include: the development of an application to provide access to the semantically structured data about medical doctors and health facilities and the creation of a repository with data obtained from municipalities, in order to allow the interoperability of information between medical and government institutions and the data management to support the decision-making.

References

Ashburner, M. et al. (s.d.), Open Biological and Biomedical Ontologies (OBO) Foundry, http://www.obofoundry.org, February 2015.

Bergman, M. K. & Giasson, F. (s.d.), UMBEL (UMBEL Vocabulary and Reference Concept Ontology), http://umbel.org, February 2015.

Bizer, C., Heath, T. & Berners-Lee, T. (2009), "Linked data-the story so far". International Journal on Semantic Web and, 5(3), pp. 1-22.

Brochhausen, M. et al. (s.d.), Ontology for Medically Relevant Social Entities, https://github.com/ufbmi/omrse, April 2015.

Guarino, N. (1997), "Understanding, building and using ontologies", International Journal of Human-Computer Studies, v. 46, p. 293–310.

Heath, T. & Bizer, C. (2011), Linked Data: Evolving the Web Into a Global Data Space, 1st ed. San Rafael, California: Morgan & Claypool.

Hitzler, P. et al. (2009), "OWL 2 Web Ontology Language Primer", http://www.w3.org/TR/2009/REC-owl2-primer-20091027, August 2014.

IHTSDO (s.d.), SNOMED-CT (SNOMED Clinical Terms), http://www.ihtsdo.org/snomed-ct, August 2014.

Noy, N. F. & McGuiness, D. L. (2001), "Ontology Development 101: A Guide to Creating Your First Ontology", http://goo.gl/NvcVTS, July 2014.

Sequeda, J. (2012), "SPARQL 101", http://goo.gl/zZzubj, August 2014.

World Wide Web Consortium (W3C) (s.d.), "Linked Data", http://www.w3.org/standards/semanticweb/data, December 2014.

An ontology of organizational knowledge

Anderson Beraldo de Araújo[1], **Mateus Zitelli**[1], **Vitor Gabriel de Araújo**[1]

[1]Federal University of ABC (UFABC)
Santo André, São Paulo – Brazil

anderson.araujo@ufabc.edu.br,{zitellimateus,araujo.vitorgabriel}@gmail.com

Abstract. *One of the main open problems in knowledge engineering is to understand the nature of organizational knowledge. By using a representation of directed graphs in terms of first-order logical structures, we defined organizational knowledge as integrated relevant information about relational structures. We provide an algorithm to measure the amount of organizational knowledge obtained via a research and exhibit empirical results about simulations of this algorithm. This preliminary analysis shows that the definition proposed is a fruitful ontological analysis of knowledge management.*

1. Introduction

According to [1], Knowledge management (KM) has produced a bunch of definitions that helps us to understand organizational knowledge, the kind of knowledge that we find in organizations. Nonetheless, there is no universal approach to the different kind of definitions available. We are in need of an ontological analysis of organizational knowledge that is capable to unify the different notion of knowledge relevant to bussiness.

Indeed, organizational knowledge has been thought according to four fundamental types [3, 4]. The first one we can call the *mental view of knowledge*. According to this standpoint, knowledge is a state of mind. In the mental view to manage knowledge involves to regulate the provision of information controls and to improve individuals capacity of applying such a knowledge. The second view is the *objectual view of knowledge*. Here knowledge is an object, something that we can store and manipulate. In the objectual approach manage knowledge becomes a process of stock managing, in which we could control the offers and the demands of individuals as parts of an integrated process inside a company. To take knowledge as a procedural phenomenon of information is the third approach, which we can call the *procedural view of knowledge*. In the procedural perspective knowledge becomes a process of applying expertise, so to manage means to manage the flows of information, such as creation process, conversion techniques, circulation processes and carrying out processes. The fourth perspective is the *credential view of knowledge*. In this approach knowledge is a credential for accessing information. In this case, KM focus on how you manage the credentials to access and what you expect to retrieve, granting the content as the result of a process.

The credential view of knowledge is the standard approach that has been applied in companies nowadays [5]. KM faces knowledge as the potential of influencing actions. By doing so companies consider KM as a process of granting the right competences to the chosen individuals. The focus is to provide the specific know-how to the realization of the processes and to grant that every processes has its correspond knowledge unit correlated.

In this paper we provide an logical method to quantify knowledge that can be used in all the four views of organizational knowledge and present computational results about them all. Quantitative indicators of knowledge can create benefits such as decreasing operational cost, product cycle time and production time while increasing productivity, market share, shareholder equity and patent income. They can drive decisions to invest on employees skills, quality strategies, and define better core business processes. Moreover, if applied to the customers, quantitative indicators can create an innovative communication platform, where the information of the clients can be quickly collected and processed into relevant decision indicators in specific terms such as abandoning one line of product, on the one hand, and investing, on the other [6, 7].

One way to unify this different approaches to KM is to outline a minimal ontology of business processes, in a Quinean sense. According to Quine, as it is well known, "to be is to be the value of a variable" [8]. In other words, ontology is the collection of entities admitted by a theory that is committed to their existence. In the present context, we call minimal ontology the ontology shared by every theory that successfully describes a processes as a organizational one. Our fundamental idea is to define organizational structures, using the general concept of first-order logical structure (Section 2). Thus, we propose a mathematical definition of information about organizational structures, based on the abstract notion of information introduced here for the first time (Section 3). The next step is to conceive organizational knowledge as justified relevant information about organizational structures (Section 4). From this approach we formulate an algorithm and simulate them (Section 5).

2. Organizational structures

We begin by some usual definitions in logic - more details can be found in [9]. The first one is associated to the syntax of organizational structures.

Definition 2.1. *A* signature *is a set of symbols $S = C \cup P \cup R$ such that $C = \{c_1, \ldots, c_k\}$ is a set of* constants*, $P = \{P_1, \ldots, P_m\}$ is a set of* property symbols*, $R = \{R_1, \ldots, R_n\}$ is a set of* relation symbols*. A* formula *over S is recursively defined in the following way:*

1. *If $\tau, \sigma \in C$, $\rho \in P$, $\delta \in R$, then $\rho\tau$ and $\delta\tau\sigma$ are formulas, called* predicative formulas*;*
2. *If ϕ and ψ are formulas, then $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$ and $\phi \leftrightarrow \psi$ are formulas, called* propositional formulas.

A theory *over S is just a set of formulas.*

Now we recall the general notion of first-order structure.

Definition 2.2. *Given a signature S, a* structure *A over S is compounded of:*

1. *A non-empty set $dom(A)$, called the* domain *of A;*
2. *For each constant τ in S, an element τ^A in $dom(A)$;*
3. *For each property symbol ρ in S, a subset ρ^A of $dom(A)$.*
4. *For each relation symbol δ in S, a binary relation δ^A on $dom(A)$.*

We write $A(\phi) = 1$ and $A(\phi) = 0$ to indicate, respectively, that the formula ϕ is true, false, in the structure A. Besides, we have the usual definitions of the logical operators $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$ and $\phi \leftrightarrow \psi$ on a structure A. In particular, we say that a theory T is *correct* about A if $A(\phi) = 1$ for all $\phi \in T$.

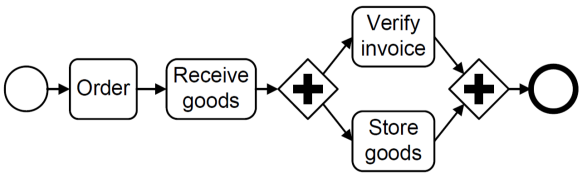

Figure 1. Bussiness process

Definition 2.3. *Given a signature S, an* organizational structure A^T *over S is compounded of a structure A over S, a theory T over S which is correct about A and expresses facts about a business process.*

The idea inside the definition of organizational structures is that they are just logical structures with a fundamental theory about how the processes works. In the proposition 2.1, we show that business processes are indeed special cases of organizational structures.

Proposition 2.1. *Business processes are organizational structures.*

Proof. According to [10], a business processes is a tuple (N, E, κ, λ), in which:

1. N is the set of nodes;
2. $E \subseteq N \times N$ is the set of edges;
3. $\kappa : N \to T$ is a function that maps nodes to types T;
4. $\lambda : N \to L$ is a function that maps nodes to labels L.

Let $L = \{l_1, \ldots, l_k\}$ be the set of labels and $T = \{T_1, \ldots, T_n\}$ be the set of types. Thus, we can define the organization structure A with domain $dom(A) = \{l_i : 1 \leq i \leq k\}$, subsets $T_1, \ldots, T_n$ of L and the relation E. □

In what follows, we write "α" $= \beta$ to mean that the symbol β is a formal representation of the expression α. Besides, X_β is the interpretation of β in the structure A, where X is a set over the domain of A. The elements of the domain $dom(A)$ of a structure A are indicated by bars above letters.

Example 2.1. *Let $S = C \cup P \cup R$ be the signature such that $C = \{i, f, o, r, s, v\}$, $P = \{E, A\}$ and $R = \{L\}$, in which "initial" $= i$, "final" $= f$, "order" $= o$, "receive goods" $= r$, "store goods" $= s$, "verify invoice" $= v$, "is event" $= E$, "is activity" $= A$ and "is linked to" $= L$. In this case, we can define the organizational structure A^T over S below, where $T = \oslash$:*

1. $dom(A) = \{\bar{i}, \bar{f}, \bar{o}, \bar{r}, \bar{s}, \bar{v}\}$;
2. $i^A = \bar{i}$, $f^A = \bar{f}$, $o^A = \bar{o}$, $r^A = \bar{r}$, $s^A = \bar{s}$, $v^A = \bar{v}$;
3. $E^A = \{\bar{i}, \bar{f}\}$ *and* $A^A = \{\bar{o}, \bar{r}, \bar{s}, \bar{v}\}$;
4. $L^A = \{(\bar{i}, \bar{o}), (\bar{o}, \bar{r}), (\bar{r}, \bar{s}), (\bar{r}, \bar{v}), (\bar{s}, \bar{f}), (\bar{v}, \bar{f})\}$.

The organizational structure A^T defined above represents the bussiness process in Figure 1.

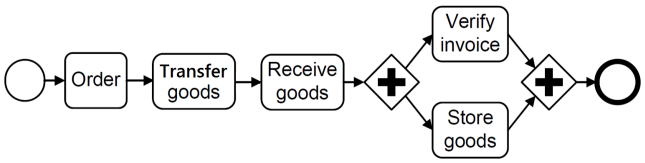

Figure 2. Extended bussiness process

3. Structural information

We turn now to the fundamental notion associated to knowledge, namely, information. The concept of information is polysemantic [11]. In this work we think of information in semantic terms. Since we are going to define a notion of information about organizational structures, we will call it *structural information*. Roughly speaking, the structural information of an organizational structure is the set of insertions and extractions that we need to perform in order to create this structure.

Definition 3.1. *Let A^T be an organizational structure over S. An* insertion *of the symbol ω into A^T is an organizational structure A_i^T such that A_i^T is an structure over $S' = S \cup \{\omega\}$ with the following properties:*

1. *$A_i^T(\tau) = A(\tau)$ for all $\tau \neq \omega$ such that $\tau \in S$;*
2. *If ω is a constant in S, then $dom(A_i^T) = dom(A)$ and $A_i^T(\omega) \neq A(\omega)$, but if ω is a constant not in S, then $dom(A_i^T) = dom(A) \cup \{a\}$ and $A_i^T(\omega) = a$;*
3. *If ω is a property symbol, then $dom(A_i^T) = dom(A) \cup \{a\}$ and $A_i^T(\omega) = A(\omega) \cup \{a\}$;*
4. *If ω is a relation symbol, then $dom(A_i^T) = dom(A) \cup \{a_1, a_2\}$ and $A_i^T(\omega) = A(\omega) \cup \{(a_1, a_2)\}$.*

Example 3.1. *Consider the organizational structure A^T over S from example 2.1. Define the signature $S' = C' \cup P' \cup R'$ equals to S except by the fact that $C' = C \cup \{t\}$, where "transfer goods" $= t$. Thus, the organizational structure A_i^T defined below is an insertion of t into A^T:*

1. *$dom(A_i) = dom(A) \cup \{\bar{t}\}$;*
2. *$t^A = \bar{t}$ and $\tau^{A_i} = \tau^A$ for $\tau \in C$;*
3. *$E^{A_i} = E^A$ and $A^{A_i} = A^A \cup \{\bar{t}\}$;*
4. *$L^{A_i} = L^A - \{(\bar{o}, \bar{r})\} \cup \{(\bar{o}, \bar{t}), (\bar{t}, \bar{r})\}$.*

The organizational structure A_i^T represents the business process in Figure 2.

Definition 3.2. *Let A be an S-structure. An element $a \in dom(A)$ is called* free *for the symbol $\omega \in S$ if there is no constant $\tau \in S$ with $A(\tau) = A(\omega)$ neither a property symbol α such that $a \in A(\alpha)$ and $a = A(\omega)$ nor a relation symbol β such that $(a_1, a_2) \in A(\beta)$ and $a_i = A(\omega)$ for $i \in \{1, 2\}$.*

If δ is a relation symbol, we write $A(\delta)_i$ to denote element a_i of $(a_1, a_2) \in A(\delta)$.

Definition 3.3. *Let A_T be an organization structure over S. An* extraction *of the symbol ω from A^T is a database A_e^T such that A_e^T is an structure over $S' = S - \{\omega\}$ with the following properties:*

1. *$A_e^T(\tau) = A(\tau)$ for all $\tau \neq \omega$ such that $\tau \in S'$;*

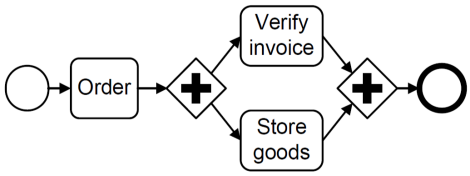

Figure 3. Contracted bussiness process

2. *If ω is a constant not in S, then $dom(A_e^T) = dom(A)$, but if ω is a constant in S, $dom(A_e^T) = dom(A) - \{A(\omega)\}$ in the case of $A(\omega)$ being free for ω, otherwise, $dom(A_e^T) = dom(A)$;*
3. *If ω is a property symbol not in S, then $dom(A_e^T) = dom(A)$, but if ω is a property symbol in S, then $dom(A_e^T) = dom(A) - \{A(\omega)\}$, where $A(\omega)$ is an element free for ω, and $A_e^T(\omega) = A(\omega) - \{A(\omega)\}$;*
4. *If ω is a relational symbol not in S, then $dom(A_e^T) = dom(A)$, but if ω is a relational symbol in S, then $dom(A_e^T) = dom(A) - \{A(\omega)_1, A(\omega)_2\}$, where $A(\omega)_i$ is an element free for ω, and $A_e^T(\omega) = A(\omega) - \{(A(\omega)_1, A(\omega)_2)\}$.*

Example 3.2. *Consider the organizational structure A^T over S from example 2.1. Define the signature $S'' = C'' \cup P'' \cup R''$ equals to S except by the fact that $C' = C - \{r\}$. Thus, the organizational structure A_e^T defined below is an extraction of r from A^T:*

1. $dom(A_e) = dom(A) - \{\bar{r}\}$;
2. $\tau^{A_e} = \tau^A$ *for* $\tau \in C''$;
3. $E^{A_e} = E^A$ *and* $A^{A_e} = A^A - \{\bar{r}\}$;
4. $L^{A_e} = L^A - \{(\bar{o}, \bar{r})\} \cup \{(\bar{o}, \bar{s}), (\bar{o}, \bar{v})\}$.

The organizational structure A_T^e represents the business process in Figure 3.

Strictly speaking, the organizational structure A_e^T in example 3.2 is *not* an extraction from A^T. For example, $L^{A^e} = L^A - \{(\bar{o}, \bar{r})\} \cup \{(\bar{o}, \bar{s}), (\bar{o}, \bar{v})\}$, which means that L^{A^e} was made of insertions in A_T as well. Since we are interested here in practical applications, we will not enter in such a subtle detail - we delegate that to a future mathematically oriented article. This point is important because it shows that to build new organizational structures from a given one is, in general, a process that use many steps. We explore this idea to define a notion of structural information.

Definition 3.4. *An* update *U^A of an organizational structure A^T over S is a finite sequence $U^A = (A_j^T : 0 \leq j \leq n)$ such that $A_0^T = A^T$ and each A_{j+1}^T is an insertion into or an extraction from A_j^T. An update $U^A = (A_j^T : 0 \leq j \leq n)$ is* satisfactory *for a formula ϕ if, and only if, either $A_n(\phi) = 1$ or $A_n(\phi) = 0$. In the case of a satisfactory update U^A for ϕ, we write $U^A(\phi) = 1$ to denote that $A_n(\phi) = 1$ and $U^A(\phi) = 0$ to designate that $A_n(\phi) = 0$. A* recipient *over in organizational structure A^T for a formula ϕ is a non-empty collection of updates $\mathcal{U}$ of A^T satisfactory for ϕ.*

Example 3.3. *Given the organizational structures A^T, A_i^T and A_e^T from the previous examples. The sequences (A^T, A_i^T) and (A^T, A_e^T) are updates of A^T that generates, respectively, the business processes in Figures 2 and 3.*

Definition 3.5. *Given a recipient $\mathcal{U}$ over a fixed organizational structure A^T, the* (structural) information *of a sentence ϕ is the set*

$$I_{\mathcal{U}}(\phi) = \{U^A \in \mathcal{U} : U^A(\phi) = 1\}.$$

Besides that, for a finite set of sentences $\Gamma = \{\phi_0, \phi_1, \ldots, \phi_n\}$*, the* (structural) information *of* Γ *is the set*

$$I_{\mathcal{U}}(\Gamma) = \bigcup_{i=0}^{n} I_{\mathcal{U}}(\phi_i).$$

Example 3.4. *Consider the recipient* $\mathcal{U} = \{(A^T, A_i^T), (A^T, A_2^T)\}$*. In this case, we have the following:*

1. $I_{\mathcal{U}}(Lrs \vee Lrv) = \{(A^T, A_i^T)\}$*;*
2. $I_{\mathcal{U}}(Lio) = \mathcal{U}$*.*

4. Organizational knowledge

Since we have a precise definition of information about organizational structures, we can now define mathematically what is organizational knowledge. The intuition behind our formal definition is that knowledge is information plus something else [12]. To be specific, we defined organizational knowledge as justified relevant information about organizational structures.

Definition 4.1. *Given an organizational structure* A^T *over* $S = C \cup P \cup R$ *such that* $C = \{c_1, \ldots, c_k\}$*,* $P = \{P_1, \ldots, P_m\}$*, and* $R = \{R_1, \ldots, R_n\}$*, the* organizational graph *associated to* A^T *is the multi-graph* $G_A = (V, \{E_l\}_{l<n})$ *such that:*

1. $V = \{(a, P_j^A) \in dom(A) \times \wp(dom(A)) : A(P_j(a)) = 1\}$ *for* $1 \leq j \leq m$*;*
2. $E_l = \{(b, d) \in V^2 : b = (a, P_j^A)) \in V, d = (c, P_k^A)) \in V, A(R_l(a, c)) = 1\}$ *for* $1 \leq l \leq n$*.*

Example 4.1. *Let* A^T *be the organizational structure from example 2.1. The organizational graph associated to* A^T *is graph* $G_A = (V, E)$ *such that:*

1. $V = \{(\bar{i}, E^A), (\bar{f}, E^A), (\bar{o}, A^A), (\bar{r}, E^A), (\bar{s}, E^A), (\bar{v}, E^A)\}$*;*
2. $E = \{((\bar{i}, E^A), (\bar{o}, A^A)), ((\bar{o}, A^A), (\bar{r}, E^A)), ((\bar{r}, E^A), (\bar{s}, E^A)), ((\bar{r}, E^A), (\bar{v}, E^A)), ((\bar{s}, E^A), (\bar{f}, E^A)), ((\bar{v}, E^A), (\bar{f}, E^A))\}$.

Definition 4.2. *Let* $\mathbb{R}^+$ *be set of non-negative real numbers. Given an organizational graph* $G = (V, \{E_i\}_{i<n})$ *associated to an organizational structure* A^T *over* S*, an* objectual relevancy *is a function* $d : V \to \mathbb{R}^+$ *and a* relational relevance *is a function* $D : \{E_i\}_{i<n} \to \mathbb{R}^+$ *such that*

$$d(a) \leq [d]$$

and

$$D(E_i) \leq [D],$$

for all $a \in V$ *and* $i < n$*.*

The functions d and D represent the relevancy associated, respectively, to the nodes and types of edges between nodes. Given that, we provide some axioms for functions that every measure of organizational knowledge must satisfy.

Definition 4.3. *We write* $U_A(G)$ *to indicate an update* $U^A = (A_j^T : 0 \leq j \leq n)$ *such that* $A_0^T = A$ *and* $A_n^T = G$*. In special,* $\mathcal{U}_A(G)$ *denotes the set of all updates* $U_A(G)$*. In this way, we define that* $K : \mathcal{U}(G_b) \times \mathcal{U}(G_r) \to \mathbb{R}^+$ *is an* knowledge function *if, and only if:*

1. $K(U(G_b), U(G_r)) = K(U(G_r), U(G_b))$;
2. *If* $G_b = G_r$ *then* $K(U(G_b), U(G_r)) = 0$;
3. *If* $G_b \cap G_r = \oslash$ *then* $K(U(G_b), U(G_r)) = 1$;
4. *If* $G_b \subseteq G$ *then* $K(U(G_b), U(G_r)) \leq K(U(G), U(G_r))$;
5. *If* $G_r \subseteq G$ *then* $K(U(G_b), U(G_r)) \leq K(U(G_b), U(G))$.

The first axiom expresses the symmetry between the knowledge base and the research base. This is a consequence of the fact that insertions and extractions are dual operations and so it does not matter whether we consider the order of the structures. The second and third axioms are immediate and the forth and fifth represent the monotonicity of the structural information.

Definition 4.4. *Let* A^T *be an organizational structure and* K *a knowledge function over an organizational graph* $G_b = (V_b, \{E_i\}_{i<n})$ *associated to* A^T*, called* knowledge base*, and an organizational graph* $G_r = (V_r, \{E_j\}_{j<n})$ *associated to an organizational structure* B^T*, called* research base*. Thus, the* organizational knowledge *of* B^T *with respect to* A^T *and* K *is the number* k *such that*

$$\mathcal{K} = \min\{K(U_{G_b \cap G_r}(G_b), U_{G_b \cap G_r}(G_r)) :$$

$$U_{G_b \cap G_r} \in \mathcal{U}(G_b) \cup \mathcal{U}(G_r)\}.$$

5. Computational results

Our approach permits us to define the algorithm *Organizational knowledge* that calculates organizational knowledge. We could provide a mathematical proof that this algorithm computes an knowledge function, but we prefer to present empirical data about its execution - in a mathematical oriented article we will give all the details. The simulations provided in this section were implemented in a program wrote in Python.

The figure Fig. 4 is a graphic $K \times |V|$, where $|V|$ is the number of nodes of a graph $G = (V, \{E_j\}_{j<n})$, generated with a number of nodes from 1 to 100 with step of 5 nodes, 5 types of edges with 10 possible values, i.e., with $n = 5$ and $D : \{E_j\}_{j<n} \to \mathbb{R}^+$ with 10 possibles values. Each knowledge measure is a result of the mean of 10 trials. This graph shows that the variation in an research base with respect to nodes are irrelevant to knowledge. This is in accordance with axiom 3. As we randomly choose new organizational graphs bigger and bigger, the probability of finding completely different graphs increase, and so knowledge approaches to 1.

The figure Fig. 5 is a graphic $K \times |E|$, where $|E|$ is the number of edges of a graph $G = (V, \{E_i\}_{i<n})$, generated with a number of nodes from 1 to 100 with step of 5 nodes, 5 types of edges with 10 possible values. Each knowledge measure is a result of the mean of 10 trials. This graph shows that the variation in an research base with respect to edges is relevant to knowledge. This is a sigmoid function, a special case of learning curve [13]. Indeed, we have obtained the following function

$$K(x) = 1/(1 + 0.001010e^{-0.385636\sqrt{x}})^{1/0.000098}.$$

The square root $\sqrt{x}$ is just due to the factor of redundancy 2.19721208941247 generated by the fact that we have chosen the graphs randomly. This redundancy implies

Algorithm 1 Organizational Knowledge

Require: $G_A = (V_A, \{E_k\}_{k<m}), G_B = (V_B, \{E_k\}_{k<n})$
Require: $d_A : V_A \to \mathbb{R}^+, d_B : V_B \to \mathbb{R}^+$
Require: $D_A : \{E_j\}_{j<m} \to \mathbb{R}^+, D_B : \{E_k\}_{k<n} \to \mathbb{R}^+$

$N := 0$
$N_A := 0$
$N_B := 0$
$R_A := 0$
$R_B := 0$
$K := 0$
for $(x, y) \in G_A$ or $(x, y) \in G_B$ **do**
  if $(x, y) \in G_A$ and $(x, y) \in G_B$ **then**
    $N := N + 1$
  else if $(x, y) \in G_A$ **then**
    for $(x, y) \in E_j$ **do**
      $N_A := N_A + 1$
      $R_A := R_A + \frac{D_A(E_i)}{[D_A]}(\frac{d_A(x)}{2[d_A]} + \frac{d_A(y)}{2[d_A]})$
    end for
  else
    for $(x, y) \in E_k$ **do**
      $N_B := N_B + 1$
      $R_B := R_B + \frac{D_B(E_i)}{[D_B]}(\frac{d_B(x)}{2[d_B]} + \frac{d_B(y)}{2[d_B]})$
    end for
  end if
end for
$K := 1 - \frac{N}{N+N_A R_A+N_B R_B}$
return K

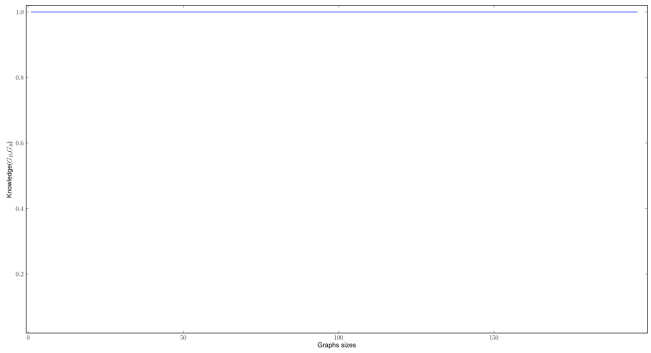

Figure 4. Knowledge between random graphs with variation of nodes

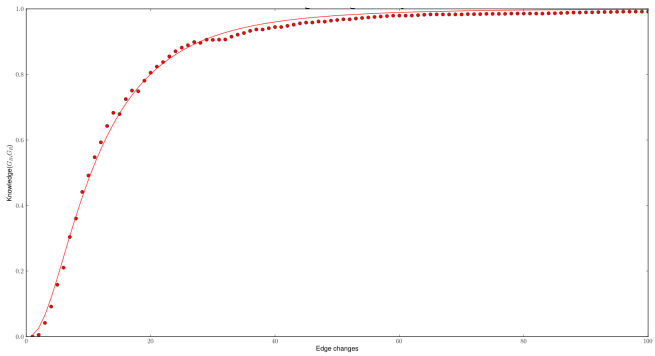

Figure 5. Knowledge between random graphs with variation of edges

a decreasing in the growing of knowledge. This is a very important result because, first, it shows a clear connection between our definition of knowledge and the usual empirical approaches to learning and, besides that, it is evidence that knowledge is indeed a relational property of organizational structures, as it have been sustained, for example, [5].

6. Conclusion

The main focus of the quantitative measure discussed in this paper is to use dynamic data taken from research methods about knowledge management. Our results shows that knowledge is a relational property of organizational structures. Nonetheless, much more should be done in order to understand the consequences of these results. At first, the organizational knowledge management techniques comprehend aspects of how to understand knowledge, using the right attitudes to the right environments. Once the knowledge meaning is defined, the knowledge sharing behaviour should be identified in order to apply quantitative measures and then driving the KM process toward a more certain path [14]. We also need to analyse how the measurement of knowledge given here can be used for these purposes. We relegate that to future works.

Acknowledgment

Omitted for the sake of double-blind evaluation.

References

[1] M. Chen and A. Chen. *Knowledge management performance evaluation: a decade review from 1995 to 2004* . [ ] Journal of Information Science, v. Feb. 13, p. 17-38, 2006.

[2] A.P. Chen and M.Y. Chen. *Integrating option model and knowledge management performance measures: an empirical study* . Journal of Information Science v. mar. 31, p. 93-381, 2005.

[3] K. Wiig. *Knowledge management: where did it come from and where will it go?* . Expert Systems with Applications, v. 13(1), p. 1-14, 1997.

[4] I. Nonaka and H. Takeuchi. *The Knowledge Creating Company* . Oxford University Press, New York, 1995.

[5] N. Boer, H. Berends and P. Van Baalen *Relational models for knowledge sharing behavior*. European Management Journal, v. 29, p. 85-97, 2011.

[6] T.H. Davenport. D.W. Long and M.C. Beers. *Successful knowledge management projects*. Sloan Management Review, v. 39(2), p. 43-57, 1998.

[7] J. Liebowitz. *Key ingredients to the success of an organization's knowledge management strategy* . Knowledge and Process Management, v. 6(1), 23-34, 1999.

[8] W. O. Quine. *From a Logical Point of View*. Cambridge, Harvard University Press, 1953.

[9] P.G. Hinman. *Fundamentals of mathematical logic*. A.K. Peters, Wellesley-Massachusetts, 2005.

[10] M. Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer, Berlin, 2007.

[11] A. Badia. *Data, Information, Knowledge: An Information Science Analysis*. Journal of the American Society for Information Science and Technology, v. 65(6), p. 1279-1287, 2014

[12] M. Zins. *Conceptual approaches for defining data, information, and knowledge*. Journal of the American Society for Information Science and Technology, v. 58(4), p. 479-493, 2007.

[13] A.C. Hax and N.S. Majluf. *Competitive cost dynamics: the experience curve*. Interfaces, v. 12(5), p. 50-61, 1982.

[14] M. Alavi and D.E. Leidner. *Review: knowledge management and knowledge management systems: conceptual foundations and research issues* . MIS Quarterly, v. 25(1), p. 107-36, 2001. ?

Ontologies in support of data mining based on associated rules: a case study in a diagnostic medicine company

Lucélia P. Branquinho[1], Maurício B. Almeida[1], Renata M.A. Baracho[1]

[1]School of Information Science, Federal University of Minas Gerais (UFMG) - Belo Horizonte - MG - Brazil

luceliabranquinho@gmail.com, mba@eci.ufmg.br, renatabaracho@eci.ufmg.br

Abstract. A well-known alternative to identify hidden standards is the use of data mining techniques. In order to obtain more efficiency in data mining, ontologies have been used to improve the representation in specialized knowledge domains. Here, we apply ontologies in a dataset of a diagnostic medicine company, which concerns to viral human hepatitis, with the aim of obtaining the best correlations between the laboratory tests prescribed by physicians and the real occurrences of diseases. Our preliminary findings show that the use of ontologies provides reduction in the number of attributes in the pre-processing phase, then improving the performance of data mining process as a whole.

1. Introduction

The amount of data stored in organizational databases has surpassed the human capacity of analysis, even considering the use of well-established technologies [Dalfovo 2000]. Thus, there is a need for adopting approaches that are able to analyze masses of data with the ultimate aim of improving the medical decision-making to both physicians and managers of healthcare organizations. A well know alternative is the approach generally referred to as Knowledge Discovery in Databases (KDD).

So, many approaches in the literature have made reference to ontologies and their semantic descriptors as a way to improve the performance of data mining based on association rules [Ferraz 2008; Vavpetic 2012; Manda 2012].

This paper aims to make an effort towards the improvement of the KDD process through the introduction of domain knowledge in the pre-processing phase. The experiment was limited to the universe of laboratory tests required for clinical analyses. In particular, we focus on diagnostics to identify viral human hepatitis. We use LOINC[1] as a reference for laboratory exam codification.

The diagnosis for viral hepatitis is based on protocols that guide the prescription of laboratory tests by doctors over the course of the disease or at an initial trial for confirmation of infections. Considering these protocols, LOINC and the research conducted in a diagnostic medicine laboratory, it was possible to map knowledge to laboratory tests viral hepatitis and reuse the OGMS[2]. In this process, we also reuse biomedical ontologies as IDO[3] , FMA[4] and DOID[5]. This mapping enabled the

[1] Available: <https://loinc.org/>.
[2] Available: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=OGMS>.
[3] Available: <http://infectiousdiseaseontology.org/page/Main_Page>.
[4] Available: <http://sig.biostr.washington.edu/projects/fm/>.
[5] Available: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology>.

generalization and the consequent reduction of the number of attributes to be mined via the identification of similarities between laboratories tests, considering the relations mapped in the viral hepatitis ontology called HVO.

The next sections will be organized as follows: section 2.1 provides a brief description on the use of ontologies in data mining; section 2.2, describes the construction of prototype of a viral hepatitis ontology; section 2.3 explains how generalization of terms will be applied in the data mining pre-processing phase as per association rules; section 3 details the results obtained with the proposed model. Finally, section 4 showcases final considerations.

2. Method

In some fields, such as Biomedicine, specialized communities have been developing and publishing, since the 1990s, a series of ontologies to aid in representing and retrieving informational [Perez-Rey et al. 2004].

2.1. Ontologies and data mining

Knowledge extraction, generally referenced in literature as Knowledge Discovery in Database (KDD) should be grouped into three phases: pre-processing, DM and post-processing. Pre-processing, which is relevant for our goals in this paper, comprises the collection, organization and treatment of data, while DM involves algorithms and techniques to search for knowledge.

Ontologies have been used to increase the relevance of the patterns discovered through the mining techniques. One of the techniques in which ontologies are being utilized is mining through association rules, which display the correlation between sets of items in series of data or transactions. [Ferraz 2008].

The advantage of pruning restrictions is to exclude information in which users are not interested in since the beginning. Every general rule should be able to replace a number of specific rules by means of generalization processes. Whenever this approach is feasible, a semantic improvement of the mined association rules and a reduction in the cardinality of the set rules will simultaneously take place.

2.2. Viral Hepatitis Ontology Construction

The development of the viral hepatitis ontology was based mapped clinical analysis laboratory tests for diagnosing human viral hepatitis considering LOINC, OGMS ontology [Scheuermann et al 2009], IDO ontology [Cowell and Smith 2006], DOID ontology [Lynn Schriml, 2009] and FMA ontology [FMA 2012]. It describes the clinical picture throughout the disease cycle by mapping terminological items that encompass diseases, their causes, their manifestations and diagnosis.

We associate the laboratory tests with the viral infectious disease to enable the generalization of the attributes to be mined, as proposed in Figure 1. In the triage phase of Hepatitis C, for example, four specific tests may be requested for the virus identification and another eight unspecific tests may be ordered for monitoring liver functions. This situation may be generalized without having denominated each laboratory tests as an attribute for data mining.

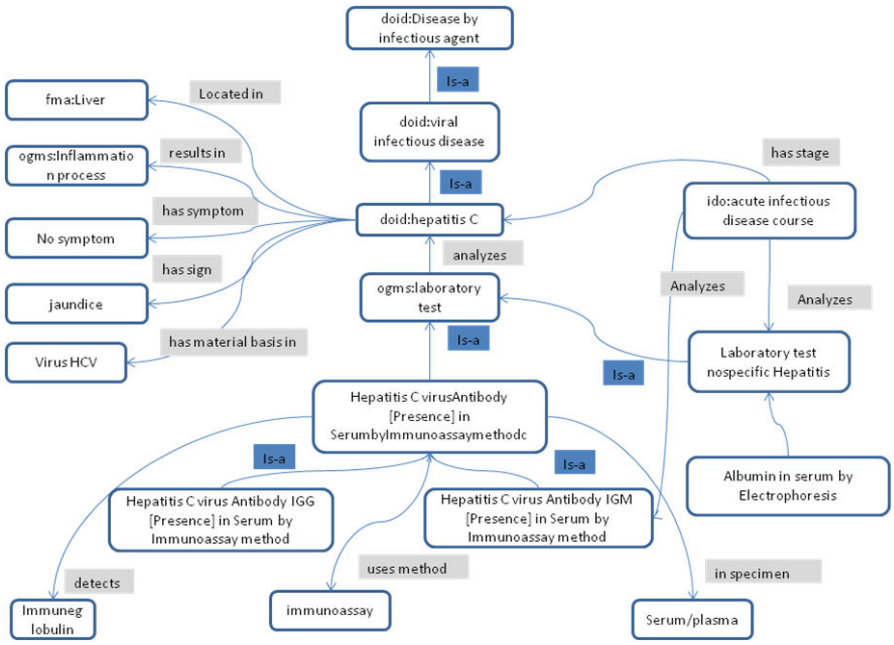

Figure 1. Example of hepatitis C classification in acute stage

The ontology for hepatitis was created at this moment of our experiment so that we could test its application in our computational architecture. We are aware that some improvements in modeling are in order, for example: i) "has symptom" and "is observed" are not ontological relations; ii) "An axioms like Hepatitis C subclass Of hasSymptom some Jaundice can be falsified by one single patient who has Hepatitis C but no jaundice"; iii) instead of "no symptom, and following OGMS, we should think in use "healthy organism". Such improvements which will be part of our future work in the following of the research.

2.3. Generalization of terms in the data mining pre-processing phase

Our study makes use of ontologies, reasoners and Jena software to promote pruning and filtering (generalization) of data from the list of laboratory tests collected from the diagnostic medicine company's database. When analyzing the relationships shared by the terms, one might identify which laboratory tests are related to which disease and stage. Therefore, the similarity between terms is considered as a means to generalize the attributes in the pre-processing phase. Figure 2 depicts the proposed model.

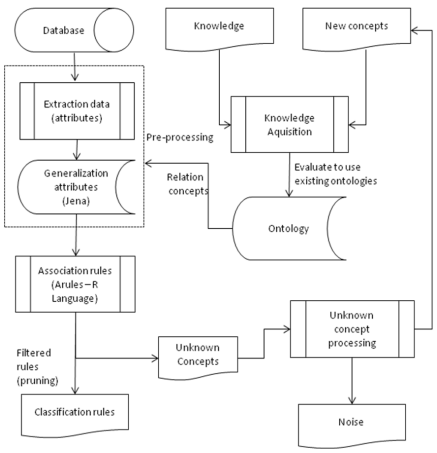

Figure 2. Model for extracting patterns of rules of association with ontologies

Based on HVO ontology, we consider relationships with the disease and with its features and also utilizing the Jena[6] tool, as well as inference rules. So, we were able to obtain more general terms to represent laboratory test groups associated with viral hepatitis diagnosis.

3. Results

The development of ontologies along with the use of inference mechanisms during the pre-mining phase has reduced the number of attributes to be mined by the association rules algorithm, namely, the Apriori algorithm. It reduced the amount of laboratory tests related to the direct diagnosis of the hepatitis virus, and also the number of unspecific tests for assessment over the course of diseases.

Based on knowledge obtained from the development of HVO, a list of laboratory test orders was collected from the company's database containing at least one test directly related to a hypothetical hepatitis virus diagnosis. Laboratory test orders that complied with the previously described rule were selected during three months, January, February and March 2015, totaling 34440 occurrences (Table 1). Test applications, which are complementary to the diagnosis, are distributed in collections made on the organization's service units, conveyed by laboratory partners throughout Brazil. In this sample, the occurrences featured 465 different laboratory tests. Considering the data of service units (support = 0.2), laboratory patterns (support = 0.02) and confidence 0.75 was executed the ARules package [Hasher 2007] in R Language to extract the association rules, we obtained the results presented below.

Table 1. Services per unit

Units	Services featuring viral hepatitis exams			Qty association rules
	Jan	Feb	Mar	Apriori
Service units	927	819	13	573
Laboratory partners	7652	7698	17331	221

Considering the same database obtained by reduction of the number of attributes, with the use of ontology, applied to 439 different laboratory tests and with the same support value and confidence was executed again the algorithm Apriori . For base units were obtained 258 and 115 rules for laboratory partners, we reached a reduction of 50% of the resulting association rules, as shown in table 2.

Table 2. Association rules generalized

Units	Services featuring viral hepatitis exams			Qty association rules
	Jan	Feb	Mar	Apriori
Service units	927	819	13	258
Laboratory partners	7652	7698	17331	115

[6] Available: < https://jena.apache.org/>.

The attributes were categorized considering the relationship between the modeled tests through equivalence axioms as showed in the example below:

hvo:laboratory_diagnostic_process_ hepatitis_A equivalent to: hvo:laboratory_testing_encounter
And (**is_composed_of** some (‘Laboratory test'
and (**diagnoses** only 'hepatitis A'))) and (**is_composed_of** min 0 ('laboratory test'
and (**diagnostic_evaluation** some Liver)))

In this equivalence axiom (described by existential restrictions), a part of the detailed diagnosis of the disease process is comprised of at least one medical application (in this case Class HVO: laboratory_testing_encounter), which is composed of complementary examinations (OGMS : laboratory test) for disease diagnostic (doid: hepatitis a) and can also be a laboratory test for evaluating the state of the health (HVO: diagnostic evaluation) of a liver (FMA: liver).

Considering the limitation of further tests and the disease is possible to identify relationships between them and, thus, promote the generalization and its representation in single attribute, in this case, a type of viral hepatitis, as shown in Table 3.

Table 3. Example of generalization

	Request – Medical prescription					Lab tests after generalization	
Patient	L.T[7]. 1	…	L.T. 8	L.T. 9	L.T. N	L.T.1	L.T. N
Patient 1	A.FETO[8]	---	TGP[9]	AU[10]	...	Hepatitis C	...
Patient 2	ALB-D[11]	---	HAV-G[12]	HAV-M[13]	...	Hepatitis A	...

Table 3 shows the laboratory tests (LT) prescribed to patients by the doctor and sent to the medical diagnostic laboratory. With the generalization attributes through the method was represented axiom disease in which some complementary tests are associated, in this case, a type of hepatitis. The other complementary tests were maintained and makes up the list of attributes analyzed by mining technique Apriori which extracted association rules related to viral hepatitis.

Therefore, our findings suggested that it is possible to reduce the rules resulting from data mining by reducing the possibilities of combining attributes. As a second, we found that the generalization of the terms enables results with a greater significance, since it can guide the post-mining phase analysis process.

4. Discussion and conclusions

The application ontology developed here strongly represents the LOINC tests for viral hepatitis, as we understand that this classification is sufficient for assessing the relationship of association rules. The extension of OGMS, DOID, FMA and IDO brings

[7]Identification of the laboratory test

[8]LOINC 1834-1 - Alpha-1 Fetoprotein – Laboratory test unspecified viral hepatitis

[9] LOINC 61151-7 - Albumin – Laboratory test unspecified viral hepatitis

[10] LOINC 5196-1 - Hepatitis B virus surface Ag

[11] LOINC 1742-6 - Alanine aminotransferase – Laboratory test unspecified viral hepatitis

[12] LOINC 5179-7 - Hepatitis A virus Ab.IgG

[13] LOINC 13950-1 - Hepatitis A virus Ab.IgM

laboratory tests closer to the diagnosis cycle and, consequently, promotes the identification of the correlations between lab tests.

It is relevant to highlight two points. Firstly, the relevance of patterns extracted by means of techniques that identify semantic similarity between terms is highly dependent of the ontology construction and validation. Therefore, it is fundamental that the domain ontologies being used be validated by specialists. Secondly, the KDD approach requires greater reach of the algorithms and cannot be restricted to “is-a” and “part-of" relations, which reinforces the use of formal semantics of ontologies.

In future work, it is intended to promote the enrichment of the ontology with new concepts and equivalence between complementary tests and disease for greater generalization of attributes.

References

Cowell, L.G, and Smith, B. (2006). Infectious Disease Ontology. In: Infectious Disease Informatics. Sintchenko V, editor. New York: Springer; 2010. pp. 373–395.

Dalfovo, O., Juarez, P., R. Alencar A., M., Palo R.D., M., J. , Otto, R., K. Silva, K. B. B. Available: <http://campeche.inf.furb.br/siic/>.

Ferraz, I. Ontology in association rules. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786067/>.

Foundational model of anatomy. Available: http://sig.biostr.washington.edu/projects/fm/.

Hasher,M., Hornik, K., Grun, B., Buchta, C., Introduction to arules – A computational environment for mining association rules and frequent item sets, 2007. Available: <http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>.

LOINC. Available: <https://loinc.org/>.

Manda, P., McCarthy, F., Bridges, S., Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships. Available: < http://www.ncbi.nlm.nih.gov/pubmed/23850840 >.

Perez-Rey, D; Maojo, V; Garcia-Remesal, M; Alonso-Calvo, R. Biomedical Ontologies.In: Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, p.207, 2004.

Piatetsky-Shapiro, G., Fayyad, U. M., Smyth, P. From data mining to knowledge discovery: an overview. Available: <http://dl.acm.org/citation.cfm?id=257942>.

Scheuermann, R. H., Werner, C., Smith, B. Toward an Ontological Treatment of Disease and Diagnosis.2009. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041577/>.

Schriml, L. (2008). Available:< http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page>

Vavpetic, A., Lavrac, N. Semantic Subgroup Discovery Systems and Workflows in the SDM. Available: <http://comjnl.oxfordjournals.org/>.

BLO: Batata Lake (Oriximiná/PA) Application Ontology

Adriano Neves de Souza, Adriana Pereira de Medeiros

Instituto de Ciência e Tecnologia – Universidade Federal Fluminense – Rio das Ostras
Rio de Janeiro – RJ – Brazil

adriano_souza@id.uff.br, adrianamedeiros@puro.uff.br

Abstract. *This work presents the BLO ontology (Batata Lake Ontology), an application ontology that describes in a structured way the data of research done by limnology researchers of Federal University of Rio de Janeiro (UFRJ) Macaé in Batata Lake (Oriximiná/PA). The main contribution of the BLO is the creation of a research data repository in RDF and the BLS application (Batata Lake System), a semantic web application to support researchers in environmental impact assessments, in preservation areas settings, in species protection and recovery of degraded areas, among other activities.*

1. Introduction

The ecological complexity of aquatic ecosystems caused by the large volume of sampling data creates difficulties to understand the environment and species, as well as the relationship between them. This understanding generates scientific data and knowledge, which provides recovery alternatives or mitigation of external impacts in the ecosystem [Bozelli et al. 2000]. Governments and organizations are encouraging solutions to share the knowledge of ecology. For example, the PELD (Long Term Ecological Program) [Esteves et al. 2004] was created by the Brazilian government to encourage the organization of research data on ecosystems. Limnology researchers of the UFRJ Macaé-RJ have been working for decades in research about the Batata Lake, an Amazonian aquatic ecosystem, located at Oriximiná-PA, that suffered environmental impacts due to the tailings generated by bauxite production [Bozelli et al. 2000]. This lake has been monitored and studied since the 80's in order to obtain knowledge of its ecosystem and mitigate these impacts. The lack of structuring and formalization of the large volume of generated data makes their analysis difficult, and limits the scope of the researchers in the search for new knowledge.

The application of Semantic Web technologies for the management and understanding of research data has been widely discussed currently. The ontologies usage in biodiversity has been appointed as a solution for obtaining scientific knowledge [Campos et al. 2011]. Ontologies for biodiversity are presented in [Moura et al. 2012], [Campos et al 2011] and [Amanqui et al 2013], but they do not describe terms proposed in this work.

This paper presents the application ontology BLO (Batata Lake Ontology) that describes the data of analysis and samplings obtained by limnology researchers of UFRJ Macaé-RJ in order to support their researches. It also presents the BLS web application for supporting the lake recovery analysis and the search for solutions that mitigate the environmental impacts. An exploratory study performed to validate the ontology is presented. Then, some conclusions and future works are discussed.

2. Batata Lake Application Ontology

Application ontologies describe concepts of a domain and specific tasks for implementing systems, the practical part [Guarino, 1997]. BLO was created following the Ontology Development 101 [Noy et al, 2001] guide. It was specified with the OWL (Ontology Web Language), specifically OWL DL 2, with 35 classes and 222 axioms. The domain was defined as Batata Lake. Thus, the ontology will be used to support the limnology researches of UFRJ Macaé, organizing research data, providing relevant information to the environmental impacts mitigation in this lake and preparing these data for online publication when needed. The ontology scope was determined by drafting the following competency questions: i) What is the sample period with the highest concentration of chlorophyll in a given year? ii) What flood pulse had the highest amount of turbidity in a given year? iii) What flood pulse had the highest percentage of organic matter in the sediment in a given year? iv) What is the flood pulse of a certain period? v) What samplings were done in impacted areas in a given period?

Searches were performed in the ontology repositories *DAML Ontology Library* (www.daml.org/ontologies/), *Protégé_Ontology_Library* (protegewiki.stanford.edu/wiki/Protege_Ontology_Library), *Schemapedia* (datahub.io/pt_BR/dataset/schemapedia) and *Swoogle* (swoogle.umbc.edu/), in order to find ontologies related to this work. The ontologies HydroBodyOfWater (sweet.jpl.nasa.gov/2.0/hydroBodyOfWater.owl) and Geography (www.daml.org/ontologies/412) contain some generic terms with descriptive features related to the proposed ontology, but they do not address the domain of this work. After the BLO definition, Albuquerque et al (2015) proposed sub-ontologies as complements to biodiversity ontology OntoBio to create a fieldwork sample vocabulary. The reuse of this vocabulary in the BLO ontology is a future work.

Figure 1 shows the graph preview of the main classes of the BLO. The vertices are classes or concepts defined in the ontology. The edges, which have a one direction, are the relations between classes, also called object properties. The *Sampling* class describes the collected sample by the researcher in the sampling stations, represented by *SamplingStation* class. *SamplingStation* has two data properties: *coordenates* and *impacted*, which respectively specify the geographical location of the sampling station and whether it is an impacted area or not. The object property *isDoneOn* determines the relation between *Sampling* and *SamplingStation*. The relation *isDoneDuring* between *Sampling* and *Period* expresses that a sampling is done in a particular period. The number of possible relations is limited by the amount of sampling stations that had some collected sample. The *FloodPulse* class specifies the lake flood pulses, which are the process stages of filling and emptying of the lake. This class has no data property, because the identification of instances is done by the URI (Flood, HighWater, ebby, LowWater). The *Period* class contains the data property *date* that describes the month and year in which the sampling is done. It is related to *FloodPulse* class by the object property *determines*. This property describes the relation between the months of the year and the flood stages of the lake, which can suffer changes over the years, because there is no standard in the establishment that a month will have a particular flood pulse. The *Sediment* and *Water* classes represent all data collected of sediment and water in the lake and they are related to the sampling by the object property *isSampliedBy*. All sampling data related to water are described by data properties of the classes *Water*, *SuspendedMatterial*, *Aluminum*, *Chorophyll*, *Iron*, *Nitrongen*, *Oxigen* and *Phosphor*.

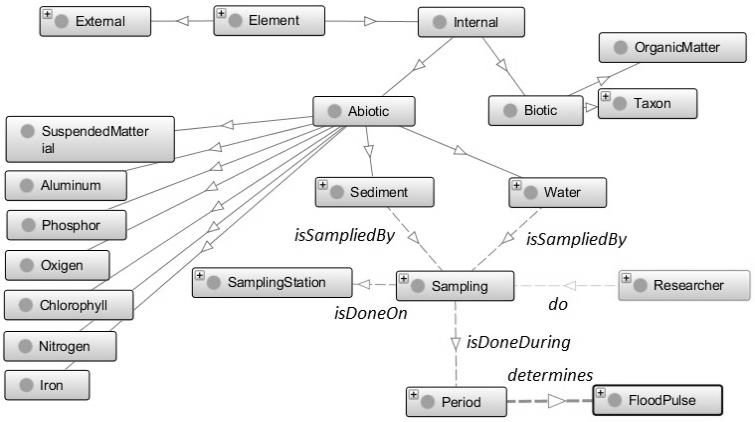

Figure 1- BLO Classes and Properties (partial)

The object property *isDoneOn* between *Sampling* and *SamplinStation* is defined with the restriction *FunctionalProperty*. Thus, a sampling *x* can be done in only one sampling station *y*. Using the triple *Sampling-> isDoneOn-> SamplingStation* is possible seek sampling information grouped by sampling stations. The object property *determines* is defined as inverse of *isDeterminedBy*. It allows that when answered the competency question "What is the flood pulse of a certain period?", the *reasoner* identifies the inverse relation *isDeterminedBy* and retrieve any instance that has the inverse as relation. Restrictions like these add semantic details to the data model and with *reasoners* the queries can obtain more accurate results, as shown in the section 3.

The BLO instances were obtained from actual research data of the Batata Lake stored in the last 26 years in spreadsheets. These data were automatically exported to RDF [Graham; Jeremy, 2004] using the BLO vocabulary and stored in a repository using the AllegroGraph 4.14 (http://franz.com/agraph/).

3. BLS Web Application

BLS (Batata Lake System) was developed to provide accurate information of the lake for researcher analysis. It was implemented in JAVA with JENA library (http://www.w3.org/2001/sw/wiki/Jena), which allows connecting the application to the RDF repository. JENA is a Java framework for building Semantic Web applications and has support for manipulating RDF triples, OWL, SPARQL [Eric; Andy, 2008] queries and includes an inference engine (*Reasoner*). The BLS interface was developed in Portuguese. Figure 2 presents the Period query page, which allows searching a given period by date (Período) or flood pulse (Pulso de Inundação). All periods of the selected pulse are raised when the page is submitted. During query performing, the application accesses the stored data in the RDF repository and run the query in SPARQL. Frame 1 presents the SPARQL query executed from page shown in Figure 2 and answers the competency question "What is the flood pulse of a certain period?". Thus, the BLS application displays the query result illustrated in the Figure 2, which shows that the flood pulse was Low Waters (AguasBaixas). Note that the data can be described using the relation *isDeterminedBy* in the RDF repository instead of *determines*. However, the query result would be the same, because these properties were defined as inverse in the BLO. The “eye” icon displays all requested period data, but the result will not be presented here due to space limitations.

Figure 2- Period Query

Figure 3 - Sampling Query

```
select distinct ?date ?floodpulse
    WHERE {
            ?period  rdfs:type  blo:Period.
            ?period  blo:determines ?floodpulse.
            ?period  blo:samplingDate  ?date.
            FILTER (?date = "12/2001")
           }
```

Frame 1 – Period SPARQL query

```
select  ?date ?station
    WHERE {
            ?period   a  blo:Period.
            ?sampling  a  blo:Sampling.
            ?station   a  blo:SamplingStation.
            ?sampling blo:isDoneOn ?station.
            ?station   blo:impacted ?impacted.
            ?sampling  blo:isDoneDuring ?period.
            ?period       blo:samplingDate  ?date.
            FILTER (?date = "12/2001").
            FILTER (?impacted = "true")
            }
  ORDER BY (?date)
```

Frame 2 – Sampling SPARQL Query

The sampling query page presented in Figure 3 allows searching the samplings done in a period or by a particular researcher in impacted area or not. It answers the competency question "Which samplings were done in impacted areas in a given period?". The application can consider the filter by researcher, otherwise it will be considered by the period. The samplings can be selected by sampling stations. The FILTER term in Frame 2 is used to determine the sampling period and the sampling station type that the researcher wants to get as answer in the sampling query page. The query result helps to evaluate the samplings which were done in impacted areas and thus comparing with samples done in non-impacted areas, in order to historically evaluate the behavior and recovery of the environment.

4. Exploratory Study

In order to evaluate the data model defined by BLO and the BLS application, a small exploratory study was conducted. The hypothesis was that the use of Semantic Web technologies for describing the Batata Lake data would facilitate the access and analysis of these data. The study was performed from a test divided into two stages: the execution of a search activity using the BLS application and the fill of an evaluation questionnaire. The activity was evaluating the water turbidity of a sample in a given period, considering as parameter the sampling data of non-impacted areas done in the same period. This is important for the researchers, since that allows evaluating the progress of the lake recovery. The study involved seven participants. The choice of them was premised on the experience and engagement with lake researches. Two of the participants, one PhD researcher and one master student, accompanied and provided all the necessary for understanding the domain and definition of competency questions.

The goal of the study was to evaluate how the research data started to be searched and analyzed using the BLS. It was not stipulated time for performing the activity. At the end of the activity each participant filled a joint questionnaire with the following questions: 1) *Do the searches available in the web application allow finding and relating the data of the samplings? Why?* 2) *Do the results obtained by the searches facilitate the comparison of the data and the analysis of the lake recovery? Why?* 3) *Would you use this application again to query and analyze your research data? Why?* 4) *Do the terms and system's menu options correspond to the everyday reality of research about the lake? If the answer is no, list the terms that do not match the reality.* 5) *Considering a scale of one to five, with option 1 equal bad and 5 equal great, how do you rate the form of searching available in the web application, comparing it with that currently performed in Excel spreadsheets?*

Most participants (five of them) answered "yes" to the questions and valued the new way to query research data. Six participants said that would use the BLS application again, as this tool significantly reduces the time spent looking for a data, enabling faster analysis. Six of them said that the vocabulary was defined according to the everyday reality of research about the lake. This indicates that the BLO ontology was well defined according to the domain. The test results also allowed identifying problems and difficulties in finding and analyzing the data. In the issue 2, the answers of four participants indicated that the queries results did not facilitate the data comparison and the lake recovery analysis, because the way the results were presented. They informed the search filter by period should be only for year interval with a flood pulse filter to facilitate analysis based on different periods and years. In addition, they suggested the choice of some variables, such as turbidity or chlorophyll, presented in parallel all the values separated by sampling stations, impacted or not. It would allow analyzing a historical series of data and effectively evaluate the lake recovery. Plus, they observed that the application navigability would be more intuitive with the access to the samplings from the data of a given period.

5. Conclusion and Future Work

This paper presented the BLO ontology for semantically describing data of the research done by limnology researchers of UFRJ-Macaé on Batata Lake. The semantic description of these data enables richer queries about the lake through inferences done by *reasoners*. In addition, it provides a vocabulary of common terms used in other researches about the Batata Lake. The main contributions of this ontology is the creation of a research data repository in RDF and the development of the BLS system, a semantic web application to support researchers to query and analysis the research data about this lake. The aim is supporting the production of scientific knowledge from the analysis made by semantic queries and preparing the data for online publication when needed. An initial exploratory study was done to validate the ontology and the application. The tests showed the BLO relevance and quality and some necessary changes in the BLS application. After implementing these changes, a new experiment will be conducted to validate them.

A future work is using the ontology proposed by Moura et al (2012) and the BLO ontology for describing the species existing in the Batata Lake. Another future work is sharing BLO so that other researchers that study this lake can use it to support their research. Moreover, some terms related to fieldwork sampling context of the

OntoBio [Albuquerque et al, 2015] can be reused.

References

ALBUQUERQUE, A. C. F., CAMPOS DOS SANTOS, J. L., DE CASTRO JÚNIOR, A. N. OntoBio: A Biodiversity Domain Ontology for Amazonian Biological Collected Objects. 48th Hawaii International Conference on System Sciences, p. 10, 2015.

AMANQUI, F. K. M.; SERIQUE, K. J.; LAMPING, F.; CAMPOS, J. L.; ALBUQUERQUE, A. C. F.; MOREIRA, D. A. Implementing an Architecture for Semantic Search Systems for Retrieving Information in Biodiversity Repositories. Simpósio Brasileiro de Banco de Dados, p. 1–6, 2013.

BOZELLI, REINALDO L.; ESTEVES, FRANCISCO A.; ROLAND, F. Lago Batata: Impacto e Recuperação de um Ecossistema Amazônico. UFRJ/SBL- RJ, 2000.

CAMPOS, J. L.; NETTO, J. F. D. M.; CASTRO, A. N. DE; ALBUQUERQUE, A. C. F. Ontologias para Interoperabilidade de Modelos e Sistemas de Informação de Biodiversidade, 2011.

ERIC, P.; ANDY, S. 2008 "SPARQL Query Language for RDF". W3C. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

ESTEVES, F. A. ; SCARANO, F. R. ; ROCHA, C. F. D. Pesquisa de Longa Duração na Restinga de Jurubatiba: Ecologia, História Natural e Conservação. 1. ed. Rio de Janeiro: RiMA Editora, 2004. v. 1. 376p.

GUARINO, N. Understanding, building and using ontologies. International Journal of Human-Computer Studies, v. 46, p. 293–310, 1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1071581996900919>.

GRAHAM, G.; JEREMY, C. 2004. "Resource Description Framework (RDF): Concepts and Abstract Syntax". W3C. Disponível em: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

MOURA, A.; PORTO, F.; POLTOSI, M. Integrating Ecological Data Using Linked Data Principles. ONTOBRAS-MOST 2012: 156-167, 2012.

NOY, N.; MCGUINNESS, D. Ontology development 101: A guide to creating your first ontology. Development, v. 32, p. 1–25, 2001. Disponível em: <http://protege.stanford.edu/publications/ontology_development/ontology101.pdf>.

A model for the construction of an inter-domain ontology: Corporate Sustainability Index and the G4 Guidelines of the Global Reporting Initiative

Tâmara Batista Reis[1], Paulo Caetano da Silva[1]

[1]Master Program in Systems and Computing -Salvador University -Brazil

tamarabr38@hotmail.com, paulo.caetano@pro.unifacs.br

***Abstract**: Ontologies are tools for knowledge representation that can help solve the diversity of representation of concepts that have similar meaning. Therefore, this paper proposes a method for building an ontology for the representation of a common knowledge base among the Corporate Sustainability Index, adopted in Brazil, with the G4 Guidelines of the Global Reporting Initiative, an international standard.*

1. Introduction

When choosing a sustainably responsible posture, an organization needs to measure, monitor and report organization's sustainability performance data, this gave rise to the concept of *Sustainable and Responsible Investment (SRI)*. Therefore, it is necessary to choose a methodology that provides the parameters through which the organization can compare its results achieved with the objectives pursued. In this sense a variety of indices and methodologies have been proposed.This diversity has caused the problem of lack of consensus and makes communication difficult between organizations that adopt different processes to manage and report their sustainability performance. Another problem concerns the lack of information standardization, because the documents requested by various stakeholders (e.g. shareholders, government) using different software, can only be obtained if there is an integration of information from heterogeneous systems. This generates costs and waste of resources for mapping this information between systems, without adding value to information.

This is the scenario in which the ontology is presented as an instrument for the representation of knowledge. This work propose a construction of an ontology, which enhances the integration the sustainability indices most used by Brazilian companies, the *Corporate Sustainability Index - ISE,* main representative of *SRI* in Brazil (BM&FBOVESPA 2014), with highly indicators worldwide, through alignment with the *Global Reporting Initiative GRI G4 Guidelines* (which are aligned to the UN Global Compact, the OECD and UNGC) wich provides a methodology of the most used worldwide [GRI 2014].The development of a taxonomy structure semantics between the relevant concepts common to the ISE / GRI will be able to provide the information quickly, efficiently, and independent methodologies. Such features can help overcome the limitations caused by the diversity indices and methodologies, and provide the integration of information, helping to overcome the computational problems faced across stakholders using heterogeneous systems.

The ontology construction activity requires the adoption of a methodology to structure the construction process [Rautenberg 2010; Luna et al., 2012]. However, by the finding of lack of consensus among the proposed methodologies for building

ontologies and the particular needs of the domains addressed, this work presents a proposal of a model for building a domain ISE / GRI ontology.

2. A proposal for the construction of the ISE/GRI ontology

The methodologies presented by Fernandez et al., (1997), and Uschold Gruninger (1996), and Noy and MacGuinnes (2001), consider ontologies like a software products and demonstrated that the development stages, are equivalent to thesoftware *life cycle phase*. These phases were adapted by the extracted processes of the IEEE-1074 standard (1997) and characteristics that are particular to ontologies, i.e. *formalization and integration*.Therefore, the IEEE-1074 (1997) was used as standard quality for the development process, describing a structured process for software development that includes all life cycle stages, described as: *project management, pre-development, development process, postdevelopment and integral processes*. Thus, based on the analyzed methodologiesand on the IEEE-1074 standard (1997), a model process has been defined, whose development phases are described as shown in Figure 1.

- **Phase 1: Project Management:** having observed the suggestion of Methontology methodology [Fernandez et al.1997], for this phase were adopted related activities beginning at the planning and project management throughout its life cycle.On the activity **Definition of the life cycle process of the ontology** is proposed that the development process is based on the evolution of prototypes [Fernandez et al. 1997].

- **Phase 2: Ontology Pre-development:** at this stage it is recommended to search the domain knowledge and the identification of problems in order to propose possible solutions through the ontology. The sources for the pursuit of knowledge can be the literature, sites, knowledge experts, etc. [Fernandez et al. 2004]. In support is recommended to perform the following activities:

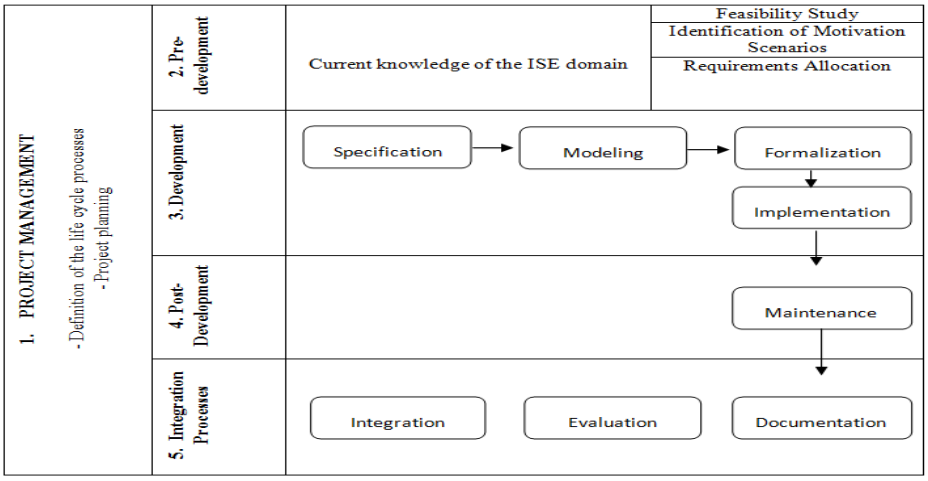

Figure 1. Phases of the development process of the Ontology

a)Feasibility Study: [ANSI/NISO Z39.19-2005]**:** is based on supplementary questions, assuggested by Silva (2008), in order to analyze the importance of building the ontology.Such questions are: 1. Why build the ontology? 2. What would happen if the ontologywas not built? 3. What are the problems with the current knowledge? 4. Howcan theproposed ontology help? 5. Will some existing ontology be reused, or be integrated? 6. Will any resources or technologies that are different from the onesalready used within the domains be needed? 7. What skills are required?

b) Identification of motivation scenarios:the motivation scenario analysis technique proposed by Gruninger and Fox (1995) helps detect the ontology domain problems and to present alternative solutions. The description of motivation scenarios are based on the identified initial requirements (which will be detailed in the requirements specification).

c) Requirements Allocation [Silva 2008]**:** this activity requires the choice of the artifacts needed to build the ontology, such as tools, *software* and *hardware*. The recommendations of this proposal for this step are: c1) for the ontology development the use of *Protégé* tool is recommended, for the following reasons : i) it has a friendly interface; ii) it documents objects; iii) it is in the public domain; iv) it has a modular architecture, which allows the inclusion of new features; v) it has a research community that contributes to its development and update; and vi) it has documentation; c2) for the ontology representation and formalization, OWL-DL is the recommended language, based on the following reasons: i) it is considered the *World WideWeb Consortium* (W3C) standard language, which enables its integration with ontologies implemented in standard Web technologies; ii) it supports axioms; iii) it provides inference mechanisms that allow to submit the ontology to evaluation; iv) it is used in *Protégé* tool, which assists in the process of implementation and formalization of the ontology; c3) for the conceptual modeling task it is suggested to use *Microsoft Visio* tool (*Microsoft*, 2014), for its usability features, user-friendly interface, and the fact that it allows the preparation of diagrams necessary for modeling the ontology.

- **Phase 3: Ontology development process:** this stage is the beginning of ontology construction process, comprising the activities described in the following.

a) Requirements specification: according to Gruninger and Fox (1995), from the observation of *Motivation Scenarios* it is possible to draw up a set of *competency questions*. These questions and their answers allow identifying information in real situations in the domain of the ontology in question. By analyzing *the questions* that the ontology will have to answer it is possible to determine the domain that the ontology should cover and delimit the ontology scope. It is recommended the documentation of this process for preparing the Scope Document Ontology, which includes information about: its purpose, its usefulness, who can use and maintain the ontology, degree of formality, responsible for the construction, sources of knowledge used, process adopted for the development, quality assurance, used tools, languages used for the representation and formalization, and the products generated.

b) Conceptual Modeling: to identify the ontology components the contribution of Silva (2008) was adopted, which reports the following elements: *conceptual classes;class attributes; instances; instance attributes; relations among the classes; constants;terms; formal axioms;* and *rules*. For best results in the *conceptual modeling* activity it is recommended to treat *the terms and concepts* involved, and only then organize them in the *taxonomic structure*. The activities flow for the conceptualization of the ISE-GRI ontology is illustrated in Figure 2.

To identify *relevant terms* it was adopted the Noy and McGuinness (2001) *proposed* which suggests questions related to *competency questions*. Such questions inquire: i) which are the terms that are relevant?;ii) what are the properties of these *terms*?; iii) what is necessary to say about these *terms*? Another contribution was taken from the ANSI/NISO Z39.19-2005 standard for the construction of controlled vocabularies, and it proposes the analysis of the domain through consultation with

several knowledge sources, according to criteria based on: i) *literary warranty* (specialized literature); ii) *structural warranty*; iii) *warranty of use*. To assist in the construction of knowledge it is suggested to use the *documentanalysis method* [Dalhberg 1978], applying a technique used in the fields of Library and Information Science, the *subject analysis technique*, recommended by Silva (2008), which assists in the identification and selection of *concepts* that represent the essence of documents,. The application of these techniques has allowed the *identification ofrelevant terms* which represent the knowledge of the ISE-GRI ontology domain, and they were recorded in the *Glossary of Terms* document, proposed by Fernandez *et al*. (1997).

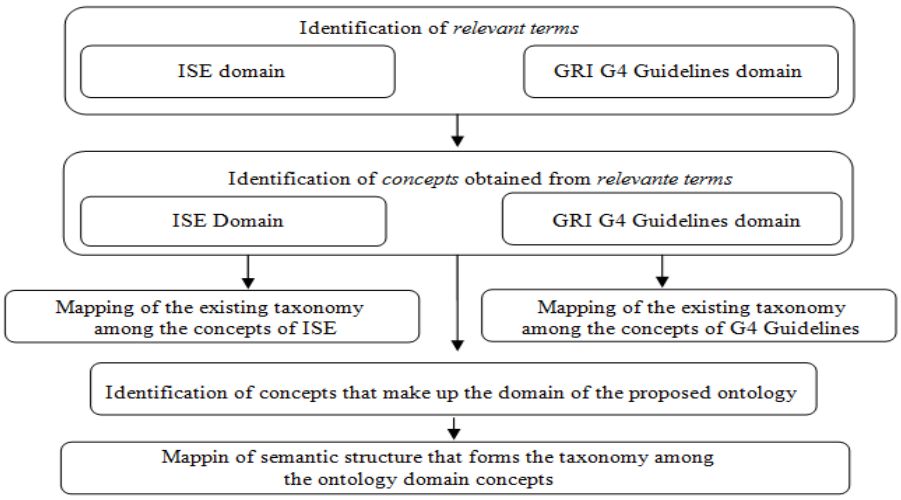

Figura 2. Fluxo de Atividades para a conceitualização da ontologia.

The next task comprises the definition of *domain concepts* and the principles adoptedare explained in the *Concept Theory* [Dalhberg 1978], which were summarized as: i) identification of the object or reference item in the domain; ii) analysis of the intrinsic and extrinsic features of the object, to define the *concept* and *relationships* among *concepts*, which allowed to form sentences about the object; iii) identification of the existing *taxonomy* among the *concepts* from the principle of contextualization, in which the definitions of *concepts* and their positions in the semantic structure are directly related to the *domain* in which the terminology is being built; iv) selection of *terms* to express the *concepts* present in the ontology.

After *defining the concepts* should be sought to know the nature of *concepts* in order to *rank these concepts into categories*, which identifies the *classes, attributes* and *relationships*. From the ANSI/NISO Z39.19-2005 standard, for the *classification of terms intocategories*, certain *facets* are determined, based on categories and subcategories of the *Concept Theory* [Dalhberg 1978], which determines the formal-categorical relationship to classify *concepts* of the same nature into a category. Thus, for each *term* the *category* to which it belongs is identified, as:*dimensions*; *activities*; *properties* and *entities*. The comparison task among the *concepts* for *classification* should consider the domain characteristics, seeking the most appropriate definition, i.e. one that meets the ontology purposes, identifying the *concepts* through the establishment of clear and unambiguous textual descriptions, defined by observing the semantic match of the meanings of terms and their relationships with each other, not in isolation or independently, as in a classical dictionary . The application of these

procedures and the elaboration of the *faceted structure* helped identify the *concepts, attributes, constants* and *relationships* of the ISE-GRI ontology domain.

The *mapping of the semantic taxonomic structure* among *domain concepts* requires analysis first ofhow the *concepts* of the same kind relate, establishing two types of relationships: a) *hierarchical*); b) *partitive*: [Dalhberg 1978]. To organize the concepts in the taxonomy and identify the levels of classes, this work suggests the combined use of methods arising from ANSI / NISO Z39.19-2005: a) top-down: identifies generic concepts, high level; b) middle-out: identifies mid-level concepts; c) bottom-up: identifies low-level concepts.As an aid in decision-making are recommended principles proposed by Noy and McGuinness (2001), which help to: a) distinguish disjoint classes; b) identifying a transitive property; c) decide by inserting / or not of new sub-classes; d) decide to create a new class or getting a property; e) decide between creating a new class or identification of an instance; f) design relations types, "is a" or "type". To ensure that the methodological process of construction of knowledge about the ontology conceptualization is correct in order to avoid distortions in the semantic meanings of the concepts, it is recommended to carry out detailed descriptions of binary relations, class attributes, the instance attributes and constant, beyond the definition of relevant concepts instances, based from models for the *intermediate representation* proposed by Fernandez et al. (2004).

c) Ontology Formalization: the formalization activity follows the suggestion by Fernandez *et al*. (2004), indicating that it can be configured through tools that generate the code (e.g. generated in the specification of axioms) by exporting the ontology specification in the representation language used by the tool. We suggested the use ofOWL-DL language in the *Protégé* tool, which is based on *descriptive logic*. This process enables the definition and formalization of *axioms* and *rules* that restrict possible deviations of domain interpretation. A formalization example of the ISE/GRI ontology is illustrated in Figure 3.

Figure 3. Existencial restriction formalization.

Figure 3 shows that *General_Standard_Disclousure* class was selected and shows the formalization of the restriction created for this class, which has the following meaning: for an individual wich is a *General_Standard_Disclousure* class member, it is necessary that this individual belongs to *G4_Guidelines_GRI* class and has at least one type of relationship with *Indicator* class, through *hasIndicator* property.

d) Implementation of Ontology: This activity aims to transform the ontology written in natural language in a computable model, capable of meeting certain requirements defined in the conceptualization phase. The terminology designed for intermediate representation models must be mapped to the constructors and axioms of OWL-DL language, Protégé tool associated with getting the same concepts, attributes, relationships and described instances. This process used for implementation of the ISE / GRI allowed the construction of ontology classes, properties, and constraints creating instances of the ontology.

➢ **Phase 4: Post-development process: t**his phase comprises the *maintenance* required to the ontology, after the completion of the *development* and *evaluation* processes, in which the necessary procedures are performed, given the identified needs [Uschold e Gruninger 1996].

➢ **Phase 5: Integration process**

a)Integration: This step includes the evaluation of high-level ontologies for the reuse of terms relevant to the conceptualization of the ontology being built.

b)Ontology Evaluation: this activity comprises technical inspections of productsthat are generated at each stage of the process, reporting the product to maintenance whenever a need for changes is detected. Otherwise, the product is documented. Gruninger and Fox (1995) suggest that the *evaluation* process to investigate the consistancy of ontology after *implementation*using the *competency questions* to observe if the ontology is able to satisfactorily respond to these questions. This paper proposes the use of OWL-DL inference engine to perform these queries to the ontology. This procedure was applied to evaluate the consistency of the ISE / GRI ontology, indicating satisfactory results, as shown in Figure 4.

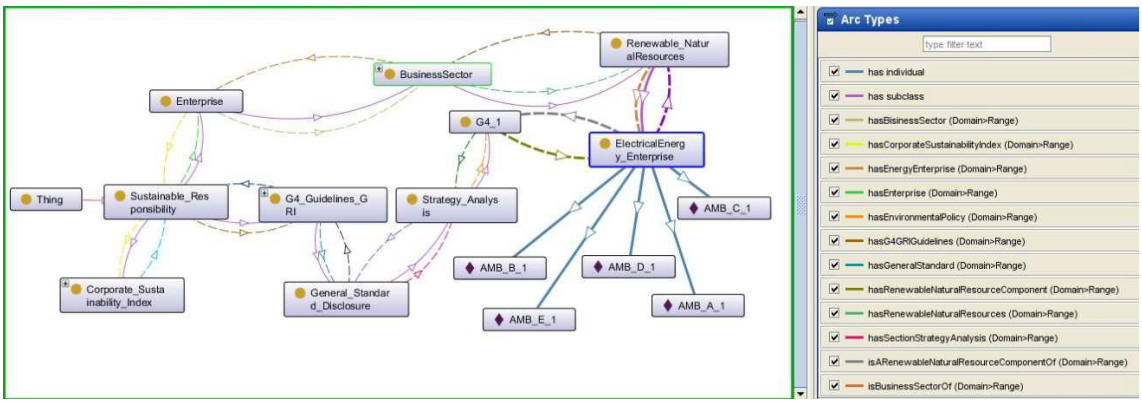

Figure 4. Evaluation of ontology in relation to *Competency Questions*

For instance, *Competency Question* inquired: Is there a relationship between environmental indicators of a company belonging to the electricity sector with the GRI indicators? "The ontology showed that a company of the electricity sector, represented by the *ElectricalEnergy_Enterprise*class, has concepts adopted by ISE (AMB_A_1 instances, AMB_B_1 , etc.) relating to the class *GRI_G4_1*.

c) Documentation: the documentation activity must be observed at all stages of the ontology *life cycle* and the generated documents should be recorded in the scope of the ontology, as suggested by Metonthology. In ISE-GRI ontology all documents were properly organized, and are available at: <http://xbrlframework.com/wiki/csa_gri/>.

3. Results obtained and future work

The results obtained by the tests performed during the *evaluation phase* showed that this proposal for the construction of the ontology attended its pre-determined purposes because the built ontology responds satisfactorily to the queries regarding the *competency questions,* as demonstrated in the evaluation section of the ontology. The proposal of this work allowed the construction of the ontology that relates the concepts of ISE with their counterparts in the G4 Guidelines of the GRI, using the construction of

a common *semantic taxonomic structure*, which represented an alignment between their indicators. This semantic environment may facilitate the manipulation of information and integration of Information Systems using these concepts.

For future work, we suggest: a) the use of this model for the construction of other domain ontologies; b) the extension of this model in more detail for the formalization and integration processes.

References

ANSI/NISO Z39.19-2005. "Guidelines for the construction, format andmanagemenet of monolingual controlled vocabularies". Bethesda: NISO Press, 2005. 176 p.

BM&FBOVESPA (2014).“Índice de Sustentabilidade Empresarial (ISE)”. Available at:<http://www.bmfbovespa.com.br/indices/ResumoIndice.aspx?Indice=ISE&Idioma=pt-BR>.

Dalhberg, I. (1978). "Teoria do Conceito". Ciência da Informação, Rio de Janeiro, v.7, n.2, p.102-107.

Fernandez, M., Gomez-Perez, O., Juristo, H. (1997). “Methontology: from Ontological Art Towards Ontological Engineering”.

Fernandez, M.; Gomez-Perez, A., Corcho, O. (2004).“Methodologies and Methods for Building Ontologies”. In: Gomez-Perez, A., Fernandez- Lopes, M, Corcho, O. Ontological Engineering. London: Springer. pp. 107-153.

GRI - *Global Reporting Iniciative* (2014)."For the Guide Lines and Standard Setting - G4". Available at: <https://www.globalreporting.org/Pages/default.aspx>.

Gruninger, M.; Fox, M. S. (1995).“Methodology for the Design and Evaluation of Ontologies”.Department of Industrial Engineering University of Toronto, Toronto, Canada M5S 1A4.

Luna, J. A. G., Bonilla, M. L., Torres, I. D. (2012). "Methodologies and methods for building ontologies." Ciência e Técnica. Año XVII, No 50, Abril de 2012. Universidade Tecnológica de Pereira. ISSN 0122-1701.

MICROSOFT (2014). Microsoft Visio - Suporte. Available at:<http://support.microsoft.com/kb/943589>.

Noy, N. F. e McGuinness, D. L. (2001).“Ontology development 101: A guide to creating your first ontology”. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Rautenberg, S., Filho, A., Todesco, J. and Ostuni-Gauthier, F. (2010) “Ferramenta ontoKEN: Uma Contribuição à Ciência da Informação para o Desenvolvimento de Ontologias”.Perspectivas em Ciência da Informação, v.15, n.1, p. 239-258.

Silva, D., (2008). “Princípios Metodológicos para Construção de Ontologias: uma abordagem interdisciplinar entre as Ciências da Informação e da Computação”. Universidade Federal de Minas Gerais.

Uschold, M. and Gruninger, M. (1996) “Ontologies: Principles, Methods an Applications”. KnowledgeEngineering Review, v. 11, n. 2.

Annotation-Based Method for Linking Local and Global Knowledge Graphs

Patrícia Cavoto[1], André Santanchè[1]

[1]Laboratory of Information Systems (LIS)
Institute of Computing (IC)
Univeristy of Campinas (UNICAMP)
Campinas – SP – Brazil

patricia.cavoto@gmail.com, santanche@ic.unicamp.br

Abstract. *In the last years, the use of data available in "global graphs" as Linked Open Data and Ontologies are increasing faster and bringing with them the popularization of the graph structure to represent information networks. One challenge, in this context, is how to link local and global knowledge graphs. This paper presents an approach to address this problem through an annotation-based method to link a local graph database to global graphs. Different from related work, the local graph is not derived from a static dataset, but it is a dynamic graph database evolving along the time, containing connections (annotations) with global graphs that must stay consistent during its evolution. We applied this method over a dataset with more than 44,500 nodes, annotating them with the values found in DBpedia and GeoNames. The proposed method is an extension of our ReGraph[1] framework that bridges relational and graph databases, keeping both integrated, synchronized and in their native representations, with minimal impact in the current infrastructure.*

1. Introduction

Real-world phenomena as biological processes, social networks and information systems have been increasingly modeled as networks, where nodes can represent individuals, computers, species, proteins, etc. and links the interaction among them. Recent research are pointing graphs as the fitted structure to store this kind of data, in which the relations among data elements are as important as the elements themselves. In the biology field, there are many uses for graphs, including metabolic networks, chemical structures and genetic maps [Vicknair et al. 2010]. The challenge is how to explore the network "behind" data available in existing information systems for analysis when data is stored in formats that do not valorize such network structure.

This challenge motivated our proposition of ReGraph, a framework inspired in the OLAP approach, which creates a special local graph database designed for network-driven analyses, aligned with an existing relational database. We applied ReGraph to taxonomic data from FishBase[2] to create FishGraph [Cavoto et al. 2015].

[1] http://patricia.cavoto.com.br/regraph/
[2] http://www.fishbase.org/

In this paper, we present an automatic annotation-based method to link our local graph database to global graphs from the Semantic Web, applied to link FishGraph data with DBpedia. Our method contributes in the data quality analysis, in the enrichment of the local database and in building the Giant Global Graph.

This is a work in progress concerning how to relate data from a local graph, stored in a graph database, with global graphs. Different from related work, our local data repository is not a static set of documents or tags to be enriched, but a dynamic graph database. It annotated content evolve along the time, bringing challenges, addressed in this research, of how to manage this hybrid graph (local and global) maintaining its consistency during the evolution.

The remainder of the paper is organized as follows. Section 2 presents the related work. Section 3 details our ReGraph framework. Section 4 presents our annotation-based approach to enrich data using ontologies. Section 5 presents our conclusions and future work.

2. Related Work

There are several contexts in which annotations are related to the Semantic Web resources (LOD and ontologies). The annotations are produced manually, semi-automatically or automatically, helping the improvement of information retrieval, knowledge reuse and information exchange [Oren et al. 2006]. There are works proposing annotations over wiki pages [Oren et al. 2006] and publishing personal notes as linked data in semantic blogs [Drăgan et al. 2010].

Several initiatives focus in how to reach semantic concepts to relate them to resources. In a survey of semantic search approaches, the authors present an overview and a classification of the existing methods for searching and browsing linked data and ontologies [Mangold 2007]. In [Alm et al. 2014] the authors propose a model to extract characteristic features from semantic annotations by importing the ontology concepts and their taxonomic relationships. Another work uses taxonomic distance measures to compute relatedness of the ontological annotations [Palma et al. 2014].

The work presented in [Santos et al. 2011] propose an architecture to discover information sources through the use of semantic search techniques in a corporative metadata repository. The process begins with an initial keyword list, followed by the query reformulation process that expands this list, adding semantically related terms and creating a new query to run on semantic annotations.

In [Amanqui et al. 2013], the authors developed a semantic search application that uses semantic web key concepts for information retrieval. They have proposed an architecture for semantic search that maps concepts of the OntoBio domain ontology to a database from the National Institute for Amazonian Research (INPA), which has collections of insects, fishes, and mammals, totalizing over 16,500 species.

As mentioned before, this work differs since it introduces a graph database perspective over the locally annotated data, which dynamically evolve along the time and must stay consistent.

3. ReGraph

As mentioned before, this method is an extension feature in our ReGraph framework, which provides a bridge integrating relational and graph databases, keeping both synchronized in their native representations. In this section, we briefly explain how the ReGraph framework works and the data conversion process from a relational to a property graph database.

3.1. The ReGraph Framework

The FishBase data is stored in a relational database. Besides the existing relational database, ReGraph produces a parallel property graph database (FishGraph), to perform network analyses and to link data with Semantic Web.

Starting from a relational database, ReGraph allows mapping its data into a property graph database, generating a *mapped subgraph*. It is also possible to further create manual and automatic annotations over this data, generating an *annotation subgraph*. Both subgraphs, *mapped* and *annotation*, are connected in the graph database. ReGraph keeps relational and graph databases in their native forms and has a synchronism module that reflects in the graph database changes executed in the relational database. The graph database is focused in the analysis on the relations among data elements.

3.2. From FishBase to FishGraph using the ReGraph framework

As previously mentioned, FishGraph concerns an application of ReGraph in the FishBase information system. We have mapped the taxonomic classification of fishes from FishBase to FishGraph - see details in [Cavoto et al. 2015]. The taxonomic classification of a species includes: Kingdom, Phylum, Class, Order, Family, Genus and Species. As FishBase has only species of fishes, it does not register Kingdom and Phylum, once that all fishes belong to the same Kingdom and Phylum. This data was compared to the taxonomic classification defined in DBpedia, generating a comparison annotation type.

In order to generate a new annotation type, we have selected also the table Country, representing countries where species are found. Figure 1 shows the graph model for the taxonomic classification and country data generated in the graph database, in which we have nodes and, associated with them, their respective properties and edges connecting it to each other.

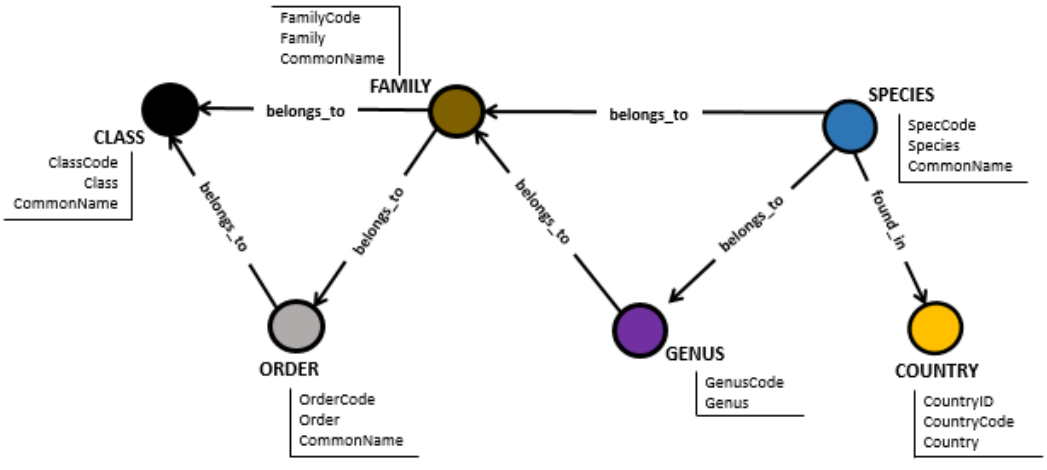

Figure 1 - Graph Model for Taxonomic Classification and Countries

We used the country information in the graph database to link them with GeoNames, a geographical knowledge base that covers all countries and contains over eight million placenames. Data retrieved from GeoNames generated new nodes and edges in the graph database, enriching it and bringing more details to the performed analyses. After the migration of the related data, we generated in the graph database 226,284 edges and 44,701 nodes, in which we have: 311 countries; 32,957 species; 10,790 genera; 572 families; 65 orders and 6 classes.

4. Automatic Annotation-Based Method

Annotations can improve the understanding and the quality of the data adding extra information. We propose a method that allows creating automatic annotations over the existent data in a property graph database. These annotations will be created through a direct connection with existing ontologies and LOD, available on the Web, e.g., GeneOntology, GeoNames and DBpedia. In this section, we detail our automatic annotation-based method and the two distinct annotation types implemented: *Comparison* and *New*. Independently of the annotation type, local data is related to Web data through a match function that compares strings to find the proper resource.

A distinctive feature of our approach is to differentiate the *annotation subgraph* (produced here) from the *mapped subgraph* (mapped from the relational database). The mapped subgraph cannot be directly changed in the graph database, since it is the product of a *one-way synchronization* originated in the relational database. Synchronization rules avoid updates in the mapped subgraph that will create inconsistencies with the *annotation subgraph*.

4.1. The Comparison Annotation Type

The main goal in the Comparison annotation type is to record comparisons of data stored in the local graph database with third party sources available on the Web. To execute this type of automatic annotation, it is necessary to define the "subject query" that will return the data from the property graph database that will be subject of the comparison.

The order of the data returned by the subject query is determinant to the correct execution of the process: (i) the first value will be the identifier of the node, helping the annotation process; (ii) the second value will be the key matched with the ontology identifiers; it will be used by the match function to retrieve data on the Web; (iii) for each of the remaining values, it is necessary specify the direct path in the ontology to reach it, linking the returned values with the specific value in the ontology; it is possible to define two paths in the ontology for each value returned by the subject query.

The result of this comparison will produce an annotation over the first node returned by the subject query. This annotation is added in the graph database as a property of the node, in which there are three possible values, annotated automatically:

- Equal: indicate elements that have the same value in the graph database and in the external ontology. This kind of annotation can improve the quality and the confidence of the data, through a double check validation.
- Not Found: represent existing elements in the graph database that was not found in the referred ontology. It can indicate: data in the graph database has spelling mistakes; the specified data does not exist in the referred ontology; data was updated in one of the sources, and was not in the other; etc.

- Divergent: represent data that have a divergence compared to the referred ontology. In can indicate: incorrect data in the graph database or in the ontology. This value is defined as a recommendation to review data. In addition, a new node is added, linked with the existing node, containing the exact data in the ontology for traceability.

4.2. The New Annotation Type

In the New annotation type, we produce new nodes, edges and/or properties, to improve the analysis and results. In this annotation type, it is necessary to specify in the "subject query" only two values: (i) the first one will be the identifier of the node, helping in the annotation process; (ii) the second one represents the key in the graph database matched with the respective identifier of a resource in the ontology; it is used by the match function to retrieve data on the Web. The second step is to define the ontology path to search.

Both data are the starting point to search in the ontology. For each information to be retrieved from the ontology and inserted in the graph database it is necessary specify: (i) ontology information: direct path in the ontology to retrieve the required information; (ii) annotation creation: how the annotation will be created in the graph database: as a node or property. The new node will be connected with the existing node by an edge that has its label also defined. In the property option, a defined property will be created over the existing node. In both cases, the value of the property will be the value found in the specified ontology.

5. Conclusions and Future Work

In this paper, we presented an automatic annotation-based method using ontologies, as an extension of our project ReGraph that connects a relational database with a property graph database, keeping both integrated, synchronized and in their native forms. It stands out for its flexibility in defining the ontologies and values that will be retrieved, compared and created, offering several possibilities to validate and enrich the graph database. Our method contrasts with the related work since it introduces a graph database perspective over the annotation-based connection between the local and global graphs. Annotations in the *annotated subgraph* stay consistent with the existing *mapped subgraph*, even after its evolution along the time.

We developed two distinct experiments to validate each proposed annotation type: Comparison and New. In the Comparison experiment, we compared almost 33,000 species of fishes from FishBase to validate their taxonomic classification with DBpedia. In the New experiment, we used the 249 countries in the graph database to retrieve their continent and information of GeonNameID and population from GeoNames.

Future work includes extending the functionality of ReGraph to allow retrieving data from other web formats and to save the link to the resource in the graph database as well as the "subject query" that generated it, helping in future repeated analysis and to track provenance.

Acknowledgments

Research partially funded by projects NAVSCALES (FAPESP 2011/52070-7), the Center for Computational Engineering and Sciences (FAPESP CEPID 2013/08293-7), CNPq-FAPESP INCT in eScience (FAPESP 2011/50761-2), INCT in Web Science (CNPq

557.128/2009-9) and individual grants from CNPq and CAPES. Thanks to FishBase.org, which provided the data used in this work.

References

Alm, R., Waltemath, D., Wolkenauer, O. and Henkel, R. (2014). Annotation-Based Feature Extraction from Sets of SBML Models. In: *Data Integration in the Life Sciences 10th International Conference, DILS 2014*, p. 81–95.

Amanqui, F. K., Serique, K. J., Lamping, F., et al. (2013). Semantic Search Architecture for Retrieving Information in Biodiversity Repositories. In: *VI Seminar on Ontology Research in Brazil, Ontobras 2013*, p. 83–93.

Berners-Lee, T. (2007). Giant global graph. Decentralized Information Group.

Bizer, C., Heath, T. and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on Semantic Web and Information Systems*, v. 5, p. 1–22.

Castilho, F. M. B. M., Granada, R. L., Vieira, R., Sander, T. and Rao, P. (2011). Ontology enrichment based on the mapping of knowledge resources for data privacy management. In: *CEUR Workshop Proceedings*, v. 776, p. 85–96.

Cavoto, P., Cardoso, V., Vignes Lebbe, R. and Santanchè, A. (2015). FishGraph: A Network-Driven Data Analysis. In: *11th IEEE International Conference on e-Science 2015*, p.1-10.

Drăgan, L., Passant, A., Handschuh, S. and Groza, T. (2010). Publishing semantic personal notes as linked data. In: *CEUR Workshop Proceedings*, v. 674, p. 1–2.

FishBase Consortium (2015). FishBase. http://www.fishbase.org, July, 2015.

Mangold, Christoph. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, v. 2, n. 1, p. 1-23.

Oren, E., Delbru, R., Möller, K., Völkel, M. and Handschuh, S. (2006). Annotation and navigation in semantic wikis?. In: *CEUR Workshop Proceedings*, v. 206, p. 58–73.

Oren, E., Möller, K. H., Scerri, S., Handschuh, S. and Sintek, M. (2006). What are Semantic Annotations?. Technical Report. http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf, July, 2015.

Palma, G., Vidal, M.-E., Raschid, L. and Thor, A. (2014). Exploiting Semantics from Ontologies and Shared Annotations to Partition Linked Data. In: *Data Integration in the Life Sciences 10th International Conference, DILS 2014*, p. 120–127.

Santos, V. Dos, Baiao, F. A. and Tanaka, A. (2011). An architecture to support information sources discovery through semantic search. In: *2011 IEEE International Conference on Information Reuse & Integration*, p. 276–282.

Vicknair, C., Macias, M., Zhao, Z., et al. (2010). A comparison of a graph database and a relational database. In: *ACM Southeast Regional Conference*, pp. 1–6.

Abordagens para Estimar Relevância de Relações Não-Taxonômicas Extraídas de Corpus de Domínio

Lucelene Lopes[1], **Maria José Finatto**[2], **Alena Ciulla**[2], **Renata Vieira**[1]

Faculdade de Informática – PUCRS
Porto Alegre – RS – Brazil

[2]Faculdade de Letras – UFRGS
Porto Alegre – RS – Brazil

{lucelene.lopes,renata.vieira}@pucrs.br {alena.ciulla,mjose.finatto}@ufrgs.br

Abstract. *This paper performs a comparison between two approaches to weight the relevance of extracted non-taxonomic relations found in domain corpora. The first approach computes the relevance according to the verb absolute frequency. The second approach computes the relevance according to the verb frequency and uniqueness in each corpus using tf-dcf relevance index, an index that takes into account the high frequency of verbs in the target corpus, and the low frequency in contrasting corpora. The preliminary results are evaluated for three domain corpora and the top relevant relations are evaluated by expert terminologists.*

Resumo. *Este artigo apresenta uma comparação entre duas abordagens de ponderação de relevância de relações não-taxonômicas extraídas de corpora de domínio. A primeira abordagem calcula a relevância de acordo com a frequência absoluta dos verbos. A segunda abordagem calcula a relevância de acordo com a frequência do verbo e sua especificidade em cada corpus utilizando o índice de relevância tf-dcf, um índice que leva em consideração a alta frequência no corpus alvo, e a baixa frequência em corpora contrastantes. Os resultados preliminares foram avaliados para três corpora de domínio e as relações mais relevantes foram avaliadas por terminologistas.*

1. Introdução

A descoberta de relações não-taxonômicas é uma tarefa difícil da aprendizagem de ontologias [Kavalec and Átek 2005]. Esta tarefa se divide em dois momentos: detectar que conceitos estão relacionados; e etiquetar cada relação detectada (em geral pela definição de um verbo que descreve a relação) [Sánchez and Moreno 2008]. Diversos trabalhos científicos seguem esse processo, por exemplo [Weichselbraun et al. 2009, Serra and Girardi 2011, Ferreira et al. 2013]. Esses trabalhos baseiam-se na detecção de verbos transitivos que relacionam dois sintagmas nominais, usualmente sujeito e objeto. Caso esses sintagmas nominais representem conceitos, ou instâncias de conceitos, esta tripla (sujeito,verbo,objeto) representará uma instância de uma relação da ontologia.

O estudo e a descrição dos verbos do português do Brasil (PB) são elementos importantes no âmbito dos Estudos da Linguagem, visto que, entre outros elementos da linguagem escrita, são elementos vitais para subsidiar uma série de recursos de representação

e de recuperação de informação com apoio computacional a partir de acervos documentais. Afinal, os verbos, além do papel fundamental para o funcionamento gramatical de qualquer língua, oferecem via predicação, elementos para a representação de diferentes relações não-taxonômicas, que extrapolam relações hierárquicas do tipo parte-todo. Os elementos relacionados assim pelos verbos podem corresponder a termos ou a conceitos de um domínio.

Um exemplo dessa importância do verbo, para a depreensão de relações entre itens de uma ontologia ou mapa conceitual é a expressão da relação não-taxonômica do tipo "X causa Y" que se depreende, por exemplo, nas seguintes frases: (1) Tabagismo causa câncer./ (2) Tabagismo pode causar câncer de pulmão./ (3) Tabagismo favorece o aparecimento de câncer de pulmão. Todavia, muitos dos trabalhos de que dispomos em PLN e mesmo em Linguística, tem se centrado na descrição de estruturas sintáticas de verbos, como a anotação de papéis semânticos ou de argumentos, que visa reconhecer padrões de associação entre determinados sujeitos e complementos nominais e um dado tipo de verbo [Scarton 2013, Zilio 2015], sem preocupação específica de caracterizar padrões de verbos em diferentes domínios.

Assumindo a existência de uma forma eficiente e eficaz de recuperar automaticamente este tipo de relação de um corpus de domínio [Lopes 2012], o desafio é filtrar dentre as relações extraídas quais são particularmente relevantes para o domínio. Na verdade, esse tipo de detecção das relações frequentemente mostra um número grande de relações e poucos estudos se dedicam a estabelecer uma ordem de relevância entre as relações detectadas.

Este artigo apresenta um trabalho inicial que propõe uma forma alternativa de estimar a relevância de relações não-taxonômicas de um domínio baseado no contraste com outros domínios. Dessa forma, a próxima seção apresenta a abordagem elementar que considera a relevância dos verbos extraídos pela sua frequência absoluta e a abordagem proposta. Em seguida, a seção 3 apresenta a comparação destas duas abordagens sobre três corpora de domínio.

2. Abordagens de Ponderação

Nesta seção apresenta-se a abordagem elementar que assume a frequência absoluta e a abordagem proposta que utiliza a frequência contrastada com outros corpora como indicador de relevância.

2.1. Abordagem por Frequência Absoluta

A primeira abordagem considerada neste trabalho toma os verbos mais frequentes como sendo as relações não-taxonômicas mais relevantes. Dessa forma, esse processo analisa o corpus alvo e identifica os verbos que relacionam dois termos, ou seja, os textos anotados são percorridos e todo verbo que conecta dois sintagmas nominais (um sujeito e um objeto) é considerado uma instância de relação. As instâncias são contabilizadas, considerando-se os verbos em sua forma canônica (infinitivo), ou seja, a frequência absoluta dos verbos é contabilizada, ignorando-se tempos verbais e flexões de pessoa e número.

A vantagem dessa abordagem é que os verbos com maior frequência absoluta serão naturalmente mais produtivos na quantidade de relações geradas, pois quanto maior

o número de instâncias, maior o número de relações a considerar. Segundo o levantamento de um grande corpus do PB [Biderman 1998], que gerou um dicionário de frequências, temos a situação sobre verbos mais frequentemente empregados, independentemente de domínio, conforme apresentado na tabela 1. No entanto, essa abordagem por frequência absoluta tenderá a privilegiar os verbos usuais semelhante aos encontrados por [Biderman 1998].

Tabela 1. Os 20 verbos lematizados no infinitivo mais frequentes no corpus de Biderman - fonte: [Finatto 2012].

ranking	verbo	ranking	verbo	ranking	verbo	ranking	verbo	ranking	verbo
1	ser	5	poder	9	dar	13	ficar	17	chegar
2	ter	6	dizer	10	ver	14	achar	18	precisar
3	ir	7	haver	11	saber	15	dever	19	começar
4	estar	8	fazer	12	querer	16	falar	20	olhar

2.2. Abordagem por Frequência Contrastada (*tf-dcf*)

Buscando ter mais especificidade nas relações a considerar, a abordagem proposta neste trabalho baseia-se na aplicação do índice *tf-dcf* (term frequency, disjoint corpora frequency) [Lopes et al. 2012]. Este índice é originalmente empregado para calcular a relevância de um termo em um corpus alvo, diretamente proporcional à frequência absoluta do termo no corpus alvo e inversamente proporcional a sua frequência em corpora contrastantes. Dessa forma, essa abordagem inicia contabilizando as ocorrências dos verbos no corpus alvo e em todos os corpora contrastantes. Em seguida, os valores de frequência absoluta dos verbos são utilizados como entrada para a fórmula do índice *tf-dcf* aplicada a verbos, em vez de termos[1] (Eq. 1).

$$\textit{tf-dcf}_v^{(c)} = \frac{\textit{tf}_v^{(c)}}{\prod_{g \in \mathcal{G}} 1 + \log\left(1 + \textit{tf}_v^{(g)}\right)} \tag{1}$$

Onde $\textit{tf}_v^{(c)}$ representa a frequência absoluta do verbo v no corpus c; e $\mathcal{G}$ representa o conjunto de corpora contrastantes.

3. Experimentos

Para ilustrar as diferenças das duas abordagens apresentadas escolhemos três corpora de domínio, um sobre Geologia (Geo), um sobre Pneumopatias (Pneumo), e o Curso de Linguística Geral (CLG), de Ferdinand de Saussure, um texto fundamental para a área de Línguistica. Adicionalmente, como a abordagem baseada no índice *tf-dcf* requer o uso de corpora contrastantes, foram utilizados três outros corpora sobre Modelagem estocástica (SM), Mineração de dados (DM) e Processamento paralelo (PP) como contrastantes. Assim, para calcular os índices *tf-dcf* de cada corpus são usados como contrastantes os dois outros corpora, além dos três corpora adicionais (SM, DM, PP). A tabela 2 apresenta as características desses corpora e indica, para os três corpora alvos o número de relações extraídas. A tabela 3 apresenta as dez relações consideradas mais relevantes para cada um dos corpora segundo a frequência absoluta (*tf*) e o índice *tf-dcf*.

Tabela 2. Características dos corpora utilizados.

corpus	Número de Textos	Número de Sentenças	Número de Tokens	Relações Extraídas
Geo	139	39,648	1,165,220	1,395
Pneumo	71	9,239	241,806	433
CLG	25	3,486	34,295	192
SM	88	44,222	1,173,401	
DM	53	42,932	1,127,816	
PP	62	40,928	1,086,771	

Tabela 3. Relações mais relevantes de cada corpus segundo ambas abordagens.

	Geo		Pneumo		CLG	
#	*tf*	*tf-dcf*	*tf*	*tf-dcf*	*tf*	*tf-dcf*
1	ser	recobrir	ser	acometer	ser	obscurecer
2	apresentar	cortar	apresentar	inalar	ter	acentuar
3	ter	aflorar	ter	contaminar	constituir	consagrar
4	mostrar	erodir	estar	contraindicar	estar	pode equiparar
5	estar	condicionar	mostrar	dever intimidar	apresentar	falsear
6	representar	retrabalhar	poder ser	poder agravar	tornar	suscitar
7	constituir	cristalizar	demonstrar	poder contaminar	fazer	unificar
8	possuir	ser depositar	revelar	poder justificar	formar	pode exprimir
9	indicar	postular	fazer	recomendar	produzir	transtornar
10	permitir	drenar	ser considerar	infectar	dar	apagar

Conforme [Biderman 1998], na sua lista dos verbos mais frequentes do PB, encabeçando-a temos os auxiliares "ser", "estar", "ter". Até o verbo "ir" registrou um elevado número de valores modais e aspectuais, razão para estar também nos primeiros lugares da hierarquia dos verbos usuais. Constam dessa lista ainda verbos modalizadores como "poder", ou vicários, e/ou suportes como "fazer", "dar"; entre os de significação plena, apenas "dizer", "falar", "olhar" e "ver" [Biderman 1998] (p. 174). Se excluirmos os verbos que integram uma locução ou que são auxiliares do levantamento por domínio com *tf-dcf* , temos que ("recobrir", "cortar" e "aflorar"); ("acometer", "inalar" e "contaminar"); e ("obscurecer", "acentuar" e "consagrar") seriam, respectivamente, os verbos de maior especificidades nos domínios de Geologia, Pneumologia e Linguística, considerando-se os corpora sob exame e os tipos de textos envolvidos.

Tabela 4. Exemplos de relações mais relevantes para o corpus Geo.

#	Frequência Absoluta	Índice *tf-dcf*
1	superfície → ser → molhável	cascalho → recobrir → formação ferruginosa
	É mostrado que a ausência de ácidos não garante que a superfície **será** molhável por a fase aquosa.	Horizonte cascalhento ferruginoso friável de superfície corresponde ao solo ou os cascalhos que eventualmente **recobrem** as formações ferruginosas.
2	footwall → apresentar → soerguimento	corpo de granito → cortar → foliação gnáissica
	Desta forma, o footwall **apresenta** sempre um soerguimento, enquanto o hangingwall é o domínio subsidente.	Os corpos de granito e pegmatito são usualmente subconcordantes, mas com freqüência **cortam** a foliação gnáissica.
3	empregado → ter → gerente	tonalito → aflorar → belt de Crixás
	Utilizando-se este paradigma, pode-se induzir que cada empregado **tem** um gerente, o que é uma generalização a partir dos dados existentes naquelas relações.	No extremo sudoeste da área o tonalito **aflora** como um corpo triangular, limitado a nordeste pelos Gnaisses Crixás Açu e a oeste pelo greenstone belt de Crixás.

[1]A única adaptação da formulação do índice *tf-dcf* para termos ao considerar verbos consiste em considerar frequência absoluta de verbos (tf_v) ao invés de frequência de termos (tf_t).

Tabela 5. Exemplos de relações mais relevantes para o corpus Pneumo.

#	Frequência Absoluta	Índice *tf-dcf*
1	efeito → ser → fator importante	espondilite tuberculosa → acometer → disco intervertebral
	O efeito idade **é** um fator importante na chance de abandono do hábito de fumar.	A espondilite tuberculosa **acomete** o disco intervertebral mais tardiamente no curso da doença.
2	moxifloxacina → apresentar → metabolização hepática	nadador → inalar → grande quantidade de ar
	A moxifloxacina, entretanto, **apresenta** metabolização hepática, e a principal via de excreção é a biliar.	Durante a prática do esporte, os nadadores **inalam** grandes quantidades de ar logo acima de a superfície da água.
3	stress → ter → papel relevante	balangeroíta → contaminar → corpos minerais
	Como exemplos, podemos citar as doenças coronarianas, em as quais o stress **tem** um papel relevante.	A balangeroíta **contamina** os corpos minerais da Itália, e assim por diante.

Tabela 6. Exemplos de relações mais relevantes para o corpus CLG.

#	Frequência Absoluta	Índice *tf-dcf*
1	língua → ser → sistema	escrita → obscurecer → visão da língua
	Visto **ser** a língua um sistema em que todos os termos são solidários e o valor de um resulta tão somente da presença simultânea de outros, segundo o esquema:	O resultado evidente de tudo isso é que a escrita **obscurece** a visão da língua.
2	língua → ter → caráter de fixidez	evolução de som → acentuar → diferença existente
	Se a língua **tem** um caráter de fixidez, não é somente porque está ligada ao peso da coletividade, mas também porque está situada no tempo.	A evolução dos sons não faz mais que **acentuar** as diferenças existentes antes de ela.
3	língua → constituir → sistema	uso → consagrar → dupla grafia
	Uma língua **constitui** um sistema.	Vimos na que, contrariamente ao que se verifica para outros sons, o uso **consagrou** para aqueles uma dupla grafia.

As tabelas 4, 5 e 6 apresentam exemplos (sentenças do corpus) das três relações mais relevantes para cada um dos corpora, respectivamente, segundo cada uma das abordagens. Observando estes exemplos, percebe-se que as relações mais relevantes segundo abordagem baseada no índice *tf-dcf* apresentam características claras de relações não-taxonômicas. Por exemplo, observa-se as triplas geradas por *tf-dcf* "cascalho **recobre** formação ferruginosa", "espondite tuberculosa **acomete** disco intervertebral", e "escrita **obscurece** visão da língua".

Já os exemplos das relações mais relevantes segundo a frequência absoluta tem um caracter que se assemelha mais a definição de propriedades/atributos, como é o caso de "superfície **é** molhável", ou ainda de "stress **tem** papel relevante". Ainda encontra-se casos que podem ser vistos como uma relação taxonômica, como por exemplo: "língua **é** sistema", ou seja, uma língua é um tipo de sistema.

4. Considerações Finais e Trabalhos Futuros

Neste estudo, mostramos dois tipos de abordagens no que diz respeito ao tratamento automático dos verbos em corpora de domínio com o propósito de identificar relações não-taxonômicas mais relevantes. Enquanto que a primeira abordagem, que considera a frequência em termos absolutos, aponta para aqueles verbos que são mais gerais da língua, a segunda abordagem, que se vale do índice *tf-dcf*, fornece uma lista de verbos que são mais específicos do domínio a que pertencem os textos.

Acreditamos, portanto, que atingimos nosso objetivo de identificar as relações mais relevantes para o domínio, contribuição do estudo através do índice *tf-dcf* que consiste no auxílio à construção de ontologias e na recuperação automática de informações, visto que acrescenta dados importantes sobre o verbo, um elemento vital - e pouco explorado, do ponto de vista do processamento automático - para o funcionamento da língua.

Além disso, temos também uma importante contribuição para os Estudos da Linguagem, ressaltando o papel dos verbos em diferentes domínios.

Cabe observar, contudo, que, quanto aos corpora em exame neste estudo, o CLG destaca-se dos outros corpora analisados, por vários motivos. Em primeiro lugar, ainda que se trate de um texto importante dentro do domínio da Linguística, não é uma compilação de textos científicos, como os corpora de Geologia e de Pneumopatias e, além disso, é uma tradução de um texto escrito originalmente em francês, em 1916. Outro aspecto é o de que é o único representante de um domínio de áreas humanas, enquanto que todos os outros são das áreas Exatas, da Saúde ou das Ciências Naturais, incluindo-se os corpora contrastantes. Por isso, fica como sugestão para trabalhos futuros, a contraposição dos verbos do CLG com os verbos de um corpus de textos de jornais, por exemplo, em que a linguagem ordinária desse gênero pode, em contraste, oferecer um panorama mais específico do domínio da Linguística.

Referências

Biderman, M. T. C. (1998). *A face quantitativa da linguagem: um dicionário de freqüências do português*. Alfa, São Paulo, Brasil.

Ferreira, V. H., Lopes, L., Vieira, R., and Finatto, M. J. B. (2013). Automatic extraction of domain specific non-taxonomic relations from portuguese corpora. In *Knowledge Discovery in Ontologies - KDO 2013*, Proc. of WI-IAT 2013, pages 161–165.

Finatto, M. J. B. (2012). Projeto porpopular, frequência de verbos em português e no jornal popular popular brasileiro. In *As Ciências do Léxico: lexicologia, lexicografia, terminologia, volume VI*, pages 277–244. Edit. da UFMS/Lab. de Edição FALE-UFMG.

Kavalec, M. and Átek, V. S. (2005). A study on automated relation labelling in ontology learning. In *Ontology Learning from Text: Methods, Evaluation and Applications*, pages 44–58. IOS Press.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Sánchez, D. and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.*, 64(3):600–623.

Scarton, C. E. (2013). Verbnet.br: construção semiautomática de um léxico verbal online e independente de domínio para o português do brasil. Master's thesis, ICMC - USP.

Serra, I. and Girardi, R. (2011). A process for extracting non-taxonomic relationships of ontologies from text. *Intelligent Information Management*, 3:119–124.

Weichselbraun, A., Wohlgenannt, G., Scharl, A., Granitzer, M., Neidhart, T., and Juffinger, A. (2009). Discovery and evaluation of non-taxonomic relations in domain ontologies. *International Journal of Metadata, Semantics and Ontologies*, 4(3):212–222.

Zilio, L. (2015). *Um Recurso Léxico com Anotação de Papéis Semânticos para o Português*. PhD thesis, PPG Letras - UFRGS.