Renata Vieira, Giancarlo Guizzardi, Sandro Rama Fiorini (Eds.)

Joint IV Seminar on Ontology Research in Brazil

and VI International Workshop on Metamodels, Ontologies and Semantic Technologies

Gramado, Brazil, September 12-14, 2011

Proceedings

http://www.inf.ufrgs.br/ontobras-most2011/

Sponsors

Organizing Institutions

Supporters

Editors' addresses:

Renata Vieira — renata.vieira@pucrs.br
Faculdade de Informática — Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Av. Ipiranga, 6681
CEP 90619-900 — Porto Alegre, RS — Brazil

Giancarlo Guizzardi — gguizzardi@inf.ufes.br
Departamento de Informática — Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, S/N, Campus Universitário de Goiabeiras, Prédio : CT-VII
CEP 29060-970 — Vitoria, ES — Brazil

Sandro Rama Fiorini — srfiorini@inf.ufrgs.br
Instituto de Informática — Universidade Federal do Rio Grande do Sul (UFRGS)
Avenida Bento Gonçalves, 9500, Caixa Postal 15064
CEP 91501-970 — Porto Alegre, RS — Brazil

Preface

Ontology is a cross-disciplinary field concerning the study of concepts and theories that support the building of shared conceptualizations of specific domains. In recent years, there has been a growing interest in the application of ontologies to solve modelling and classification problems in diverse areas such as Computer Science, Information Science, Philosophy, Artificial Intelligence, Linguistic, Knowledge Management and many others.

The Seminar on Ontology Research in Brazil, ONTOBRAS, foresees an opportunity and scientific environment in which researchers and practitioners from Information Sciences and Computer Science can discuss the theories, methodologies, languages, tools and experiences related to ontologies development and application.

In particular, this Seminar fourth edition was held simultaneously with the sixth edition of the International Workshop on Metamodels, Ontologies, Semantic Technologies (MOST), as the result of an effort of the research community in integrating the events on Ontologies which happened in Brazil in recent years. The goal was to create a unique highly scientifically qualified international forum for presenting and discussing this important topic.

The event was organized by the *Federal University of Rio Grande do Sul* (UFRGS) through the *Institute of Informatics* and *Faculty of Library and Communication Sciences*. It was also supported by the *Pontifical Catholic University of Rio Grande do Sul* (PUCRS), the *Federal University of Espírito Santo* (UFES) and *The International Association for Ontology and its Applications* (IAOA). The event was partially funded by **CAPES Foundation**, from the Brazilian Education Ministry, and by **ENDEEPER Knowledge Systems**.

Researchers and practitioners were invited to submit theoretical, technical and practical research contributions that directly or indirectly address the issues above. The call for papers was open for two categories of submissions: Full papers (maximum 12 pages) written in English and describing original work with clear demonstrated results. Accepted full paper were invited for oral presentation. The second category was short papers (maximum 6 pages), written in Portuguese, or English, or Spanish and describing ongoing work. Accepted short papers were be invited for poster presentations.

We received 29 full-paper submissions, out of which 7 are accepted for publication and oral presentation; and 36 short-paper submissions, out of which 16 are accepted for publication and poster presentations. Also, the reviewers considered that another 7 full papers, while too premature to be accepted as full papers, are relevant enough to be presented as posters. This volume is thus constituted by 7 full papers and 20 short papers, selected by our program committee, which is composed by national and international referees.

We thank the organizing committee for their commitment to the success of the event, the authors for their submissions and the program committee for their hard work.

September 2011

Renata Vieira
Giancarlo Guizzardi
Sandro Rama Fiorini

General Chairs

Mara Abel (INF - UFRGS)
Sônia Elisa Caregnato (FABICO - UFRGS)

Program Chairs

Renata Vieria (INF - PUCRS)
Giancarlo Guizzardi (INF - UFES)

Editorial Chair

Sandro Rama Fiorini (INF - UFRGS)

Organizing Committee

Carlos Alberto Silveira Jr. (INF - UFRGS)
Glória Isabel Sattamini Ferreira (FABICO - UFRGS)
Joel Carbonera (INF - UFRGS)
Leoci Sciortino (CEI - UFRGS)
Lourdes Tassinari (INF - UFRGS)
Martha Eddy K. K. Bonotto (FABICO - UFRGS)
Rafael Port da Rocha (FABICO - UFRGS)
Regina Helena van der Lan (FABICO - UFRGS)

Program Committee

Mauricio Almeida (UFMG, Brazil)
Fernanda Araujo Baiao (UNIRIO, Brazil)
Marcello P. Bax (UFMG, Brazil)
Ig Ibert Bittencourt (UFAL, Brazil)
Stefano Borgo (Laboratory for Applied Ontology [LOA], Italy)
Regina Braga (UFJF, Brazil)
Marisa Brascher (UnB, Brazil)
Ligia Café (UFSC, Brazil)
Gilberto Camara (INPE, Brazil)
Maria Luiza De Almeida Campos (UFF, Brazil)
Sônia Caregnato (UFRGS, Brazil)
Rove Chishman (Unisinos, Brazil)
Oscar Corcho (UPM, Spain)
Evandro Costa (UFAL, Brazil)
Paulo Costa (George Mason University, USA)
Alicia Diaz (LIFIA, Argentina)
Jérôme Euzenat (INRIA, Grenoble Rhône-Alpes, France)
Ricardo Falbo (UFES, Brazil)
Roberta Ferrario (Laboratory for Applied Ontology [LOA], Italy)
Frederico Fonseca (Penn State, USA)
Fred Freitas (UFPE, Brazil)
Renata Galante (UFRGS, Brazil)
José Augusto Chaves Guimarães (UNESP, Brazil)

Claudio Gutierrez (University of Chile, Chile)
Gabriela Henning (INTEC/UNL, Argentina)
Wolfgang Hesse (Philipps - University Marburg, Germany)
Seiji Isotani (University of São Paulo, Brazil)
Nair Kobashi (USP, Brazil)
Werner Kuhn (University of Muenster, Germany)
Fernanda Lima (UnB, Brazil)
Gercina A B O Lima (UFMG, Brazil)
Vânia Lima (USP, Brazil)
Maria Machado (UFRJ, Brazil)
Andreia Malucelli (PUC-PR, Brazil)
Riichiro Mizoguchi (Osaka University, Japan)
Regina Motz (Universidade de la Republica, Uruguay)
Ana Maria Moura (LNCC, Brazil)
Fernando Naufel (UFF, Brazil)
Alcione Oliveira (UFV, Brazil)
José Palazzo M. De Oliveira (UFRGS, Brazil)
Jose M Parente De Oliveira (ITA, Brazil)
Fernando Parreiras (University of Koblenz-Landau, Germany)
Paulo Pinheiro Da Silva (University of Texas at El Paso, USA)
Fabio Porto (LNCC, Brazil)
Florian Probst (SAP, Germany)
Renato Rocha Souza (FGV-RJ, Brazil)
Ana Carolina Salgado (UFPE, Brazil)
Plácida da Costa Santos (UNESP, Brazil)
Stefan Schulz (University of Graz, Austria)
Daniel Schwabe (PUC-RJ, Brazil)
Renata Vieira (PUCRS, Brazil)
Gerd Wagner (Brandenburg University of Technology at Cottbus, Germany)
Renata Wassermann (USP, Brazil)

Contents

Part I

Full Papers

A Domain Ontology to Support Evidence-Based Practice and Context Usage on Crime Prevention[1]

Expedito C. Lopes[1], Gabriela P. R Pinto[2], Vaninha Vieira[3], Teresinha F. Burnham[4]

[1]Program in Systems and Computing – Salvador University (UNIFACS)
Salvador – BA – Brazil

[2]Informatic Area – Department of Exact Sciences – State University of Feira de Santana (UEFS) – Feira de Santana - BA – Brazil

[3]Computer Science Department – Federal University of Bahia (UFBA)
Salvador – BA – Brazil

[4]Faculty of Education – Federal University of Bahia (UFBA)
Salvador - BA – Brazil

ditoexpe@gmail.com, vaninha@dcc.ufba.br, gabrielarprp@gmail.com, tfroes@ufba.br

Abstract. *Evidence-Based Practice (EBP) represents a decision-making process centered on justifications of relevant information. Context is a kind of knowledge that supports the ability to define what is or is not relevant in a given situation. The decision-making context can have an impact on evidence-based decision-making, but the integration of evidence and context is still an open issue. Ontology is referred as the shared understanding about a domain. One of the main reasons for developing context models based on ontologies is the knowledge sharing that enables computational entities, such as agents and services to find actors' similar profiles in decision making environment. This paper presents the integration of evidence and context on decision making and proposes a domain ontology to support EBP and context usage on the crime prevention domain. A practical implementation serves to validate our work.*

1. Introduction

The Evidence-Based Practice (EBP) paradigm, usually employed in several areas such as Medicine, Crime Prevention, Education and Software Engineering, are systematic procedures that take into account a problem being faced by an actor (e.g. diabetes in children), his/her needs and preferences for decision, leading to a search for evidence and an application based on the best research evidence found (Sacket et al. 2001). The procedures represent an evidence-based decision-making process, centered on justifications of relevant information (Dobrow et al. 2004).

Context is a knowledge that supports the ability to define what is or is not relevant in a given situation (Vieira et al. 2010). The application of evidence to a

[1] This work was supported by the National Institute of Science and Technology for Software Engeneering (INES+), founded by CNPq grant 573964/2008-4. http://www.ines.org.br.

particular patient, for example, detains important contextual information in the EBP procedures and includes comparative analysis between different contexts: that of the generation of evidence and that of the patient.

According to Dobrow et al. (2004) “the two fundamental components of an evidence-based decision are evidence and context. The decision-making context can have an impact on evidence-based decision-making”. There is significant research in the fields of EBP and context. However, the integration of evidence and context in computer models is still an open issue.

Ontology is referred as the shared understanding of some domains, which is often conceived as a set of entities, relations, functions, axioms and instances. One main reason for developing context models based on ontologies is knowledge sharing that enables computational entities, such as agents (human or software) to find similar profiles in decision making environment (Wang et al. 2004).

This paper aims at: (i) presenting the integration of evidence and context concepts preserving their characteristics of representation for domains that use EBP; and (ii) describing domain ontology to support the search and retrieval of evidence, regarding their contexts. This ontology is a start point to provide arguments for a semantic formulation about the characteristics of a problem, increasing the evidence-based solution, in the crime prevention domain. The motivation behind the ontology construction is due to the lack of ontologies adaptable to our purpose. Therefore, this paper also aims at providing artifacts to support system designers and provenance community experts.

The evidence retrieval increased with contextual information can also facilitate reapplying decision-making justifications when similar problems occur. The context usage also allows filtering out and sharing more useful information so the retrieved information can meet the decision maker needs. In this sense, context becomes a significant tool to optimize performance and reduce search results. Filtering mechanisms avoid more explicit user interactions with the application (Bunningen, 2004).

The remaining of the paper has the following organization. The key concepts regarding context and evidence are described in Section 2. Section 3 presents a meta-model that integrates evidence and context high level concepts. In Section 4, the domain ontology for the Crime Prevention domain, integrated with the meta-model concepts, is described. An implementation for a scenario of usage is presented in Section 5, which serves to validate our work. Related Works are described in Section 6. Finally, in Section 7 we present our conclusions and directions for further work.

2. Background

This section defines context and provides an overview of Evidence-Based Practice.

2.1. Context

There are several definitions about context. A classical definition is proposed by Dey and Abowd (2001) for whom context is “any information that characterizes the situation of an entity, where this entity is a person, place or object considered relevant in the

interaction between the user and an application. A context is typically the location, identity and status of people, groups and computational and physical objects". Context can also be seen as a set of conditions and relevant influences that make a situation unique and understandable (Brézillon 2007) or as a set of information items (e.g. concepts, rules and propositions) associated with an entity (Vieira et al. 2010).

An item is considered part of a context only if it is useful to support a problem solving. This item corresponds to a contextual element defined as "any data, information or knowledge that enables one to characterize an entity on a given domain" (Vieira et al. 2010). Contextual information regarding acquisition is: (i) given by the user, whether from persistent data sources or from profiles; (ii) obtained from a knowledge base; (iii) obtained by means of deriving mechanisms; or (iv) perceived from the environment (Henricksen and Indulska 2006). It is usually identified through the dimensions *why, who, what, where, when* and *how* (Brézillon 2007).

One step in the task execution or problem-solving process is known as *focus*. The contextual elements should have a relevant relationship to the focus of a human agent or software agent. In general, focus is what determines which contextual elements should be instantiated (Brézillon 2007).

2.2. Evidence-Based Practice

According to Thomas and Pring (2007), in general, information labeled as evidence is those whose collection has concerns about its validity, credibility and consistency with other facts or evidences. In relation to its credibility, evidences are categorized in ways:

1. Based on professional practice, as a clinical examination;
2. Generated by a process involving scientific procedures with a proven history in producing valid and reliable results, e.g a collect performed by biomedical;
3. Based from published research that corresponds to critical reviews of the area, such as randomized clinical trial.

"Evidence" in EBP, also called "research evidence", corresponds to the third category above and means a superior type of scientific research proof, such as generated through systematic review and meta-analysis in the highest level. These published researches are available in reliable data bases, usually found on sites over the Internet, carried out by independent research groups (Sackett et al. 2001). This is the concept of evidence applied in this paper.

To clarify further, a systematic review is a review that presents meticulous research and critical evaluations of primary studies (case study, cohort, case series, etc.), based on research evidence related to a specific *theme*. It contains analysis of *qualitative* results conducted in distinct locations and at different times. Meta-analysis is a systematic review of *qualitative and quantitative* characteristics (Friedland et al. 1998).

Evidence-Based Practice (EBP) involves complex decision-making, based on available research evidence and also on characteristics of the actor of the problem, his/her situations and preferences (Sackett et al. 2001).

In the medical area, EBP focus is to provide effective counseling to help patients with terminal or chronic illness to make decision in order to extend or increase the quality of their life (Friedland et al. 1998). What is objectively searched is "the

integration of best research evidence, clinical skill and patient's preferences, regarding individual risks and the benefits of proposed interventions" (Sackett et al. 2001).

In crime prevention, EBP involves the correlations practice that has been proven through scientific research, aimed at reducing the recidivism of offenders. EBP primarily considers the risk and need principle of the offender, besides the motivation, and treatment and responsibility principles (Warren 2007).

The EBP focus for education area is improving the quality of research and evaluation on education programs and practices, and hence, the information diffusion in the *educational research* field to be used by professionals and policies creators (Thomas and Pring 2004).

So, we generalize the EBP steps in the following way:

1. Transforming the need for information into a question that can be answered;
2. Identifying the best evidence to answer the question;
3. Critically analyzing the evidence to answer:
   - Is it valid (appropriate methodology and proximity to the truth)?
   - Is it relevant (size and significance of the observed effects)?
   - Can it help (applicable in professional practice)?
4. Integrating critical analysis with professional skills and the values and cultural aspects of the actor of the problem answering:
   - How much the evidence can help the actor in particular?
   - Is it adaptable to actor's goal and preferences?
   - How much safety can be expected?
5. Evaluating the efficiency and effectiveness of the results of each step for future improvement.

3. A Meta-model to Represent Evidence-Based Practice and Context Usage

The primary aim of a meta-model is to provide a set of building blocks and rules used to build models (Chomsky 1965). In this perspective we propose a class structure that represents information related to EBP procedures, while taking into consideration information about its decision-making context.

Thus, domain analysis was done in the crime prevention (particularly in juridical and social work), medical and educational environments, including: bibliographical research (Warren 2007; Satterfield et al. 2009; Sacket et al. 2001; Friendland et al. 1998; Thomas and Pring, 2004, etc.), specific legislation research, analysis of real cases collected and interviews with decision-makers from Pernambuco state court, Brazil.

Figure 1 below presents a meta-model that corresponds to integration of EBP with contextual information. We use the extension construct *stereotype* of the UML to select enumerated values. To facilitate its presentation in a systematic way, it became convenient to group classes in two integrated packages: *Context* and *Evidence*.

3.1. *Context* Package

The classes of the *context* package are based on Vieira et al. (2010). The *focus* is treated as an association of a *task* with an *agent*, which have a *role* in problem resolution. A

task "make a critical analysis of the best evidence found" for a “medical” agent in the role "evaluator", serve as example. *ContextualEntity* represents the entities of the application conceptual model and is characterized by at least one contextual element. A contextual element is a property that can be identified by a set of attributes and relationships associated with *ContextualEntity* (Vieira et al. 2010). *Accessibility* is an example of a contextual element for the *Document* class. The association between *Focus* and *ContextualElement* determines what is relevant for a focus.

Characteristics attributed to the type of context (dimension) and the method of acquiring contextual elements are considered in the framework. *Contextual sources* may be internal or external to the decision-making environment (e.g., the patient’s medical records, a document with evidence obtained from websites).

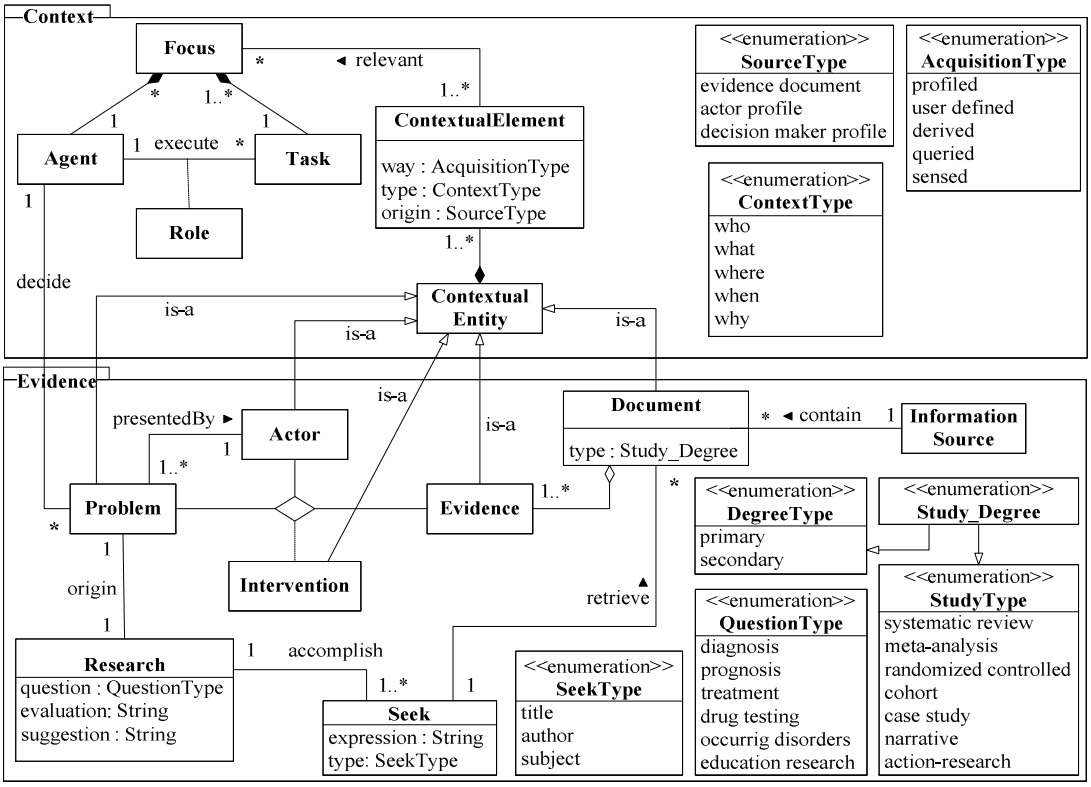


Figure 1. A meta-model integrating evidence with context

3.2. *Evidence* Package

The starting point is the observation of a problem presented by an actor to be decided by agent. Each problem is associated with an inquiry that is initiated by a formulated question (see step 1 of the EBP procedures), and completed with a self-evaluation of the research performance and suggestions for the future (see step 5 of the EBP procedures), whose information is instantiated in the *Research* class. Each domain in which EBP is applied has a list of different types of questions. For example: "diagnosis" and "prognosis" in the medical area, "drug testing" and "occurring disorders" in the area of crime prevention, and "educational research" in education.

During the evidence research, several searches can be performed to retrieve documents. For the *Seek* class, the expression and the type of search must be present.

InformationSource represents database sources that hold documents with research evidences, such as Cochrane Library, Campbell Library or Springer International Publisher, and the evidences were generated by independent research groups of specific areas (e.g. Cochrane Collaboration for medical area and Campbell Collaboration for areas of education and crime prevention).

Each document presents a type of study that can be in all domains (e.g. systematic review, case study) or in specific domain (cohort - in the medical area; action-research - in education). Systematic review and meta-analysis are studies of second degree; the remains are of first degree (Friedland et al. 2001).

After finding evidences, the agent (decision maker) will choose the one that seems the most appropriate (step 2 of EBP), which is instantiated in the *Evidence* class.

The result of the critical analysis – the validity, relevance and applicability of the best evidence (step 3 of EBP) – corresponds to contextual information. Relevance is a contextual element in *Document*, while applicability (practical utility) is in *Evidence*. Thus, *Document* and *Evidence* are specializations of *ContextualEntity*.

The *Intervention* class is the result of an association among the *Problem*, *Actor* and *Evidence* classes. It contains a description of a decision made (intervening solution) where information about associated classes have been considered including preferences, values and cultural aspects (conduct, behaviour, for example) of the actor with the problem presented (step 4 of EBP). A preference is a contextual element and hence *Actor* is a specialization of *ContextualEntity*. Problem aspects, such as the circumstances about a juridical fact for the criminal area, generally, are contextual elements used to diagnose the problem. So, the *Problem* class is a *ContextualEntity* too.

Summarizing, some elements that characterize a meta-model are shown through some examples. The *Agent* class corresponds, respectively, to the *Doctor, Judge* and *Professor* classes for the medical, juridical and educational areas. *Evidence* and *Seek*, for example, are general classes for any domains. The classes *ClinicalProblem*, in the medical domain, and *JuridicalFact*, in the juridical domain, represent the *Problem* class of the meta-model.

4. A Domain Ontology for the Crime Prevention

This section describes the main steps in the construction of ontology for representing EBP considering contextual information in the crime prevention domain. The ontology is constructed using the Web Ontology Language (OWL). To edit the ontology and axioms we used the Protégé (http://protege.stanford.edu).

4.1. Ontology Concepts

Figures 2 and 3 show a set of subclass/superclass of the main concepts defined in the crime prevention ontology. The concepts were constructed based on a survey of the concepts related found in the technical and scientific literature (Warren 2007; Gomes 2008; Moreira 2007; Saliba 2009). In this section, we present the specific concepts for the crime prevention domain, since the high level concepts concerning context and evidence were described in the meta-model (Section 3).

The ontology comprises two main classes: *ContextualEntity* and *ConventionalEntity*. *ContextualEntity* contains the subclasses that detain at least a contextual element (or contextual property), which supports the description of scenarios found in environmental decision making. According to the meta-model, described in Section 3, *ContextualEntity* has six main subclasses: *Agent*, *Actor*, *Problem*, *Document*, *Evidence* and *Intervention*. For the crime prevention domain, we defined the subclasses illustrated in Figure 2 and summarized the main contextual properties in Table 1.

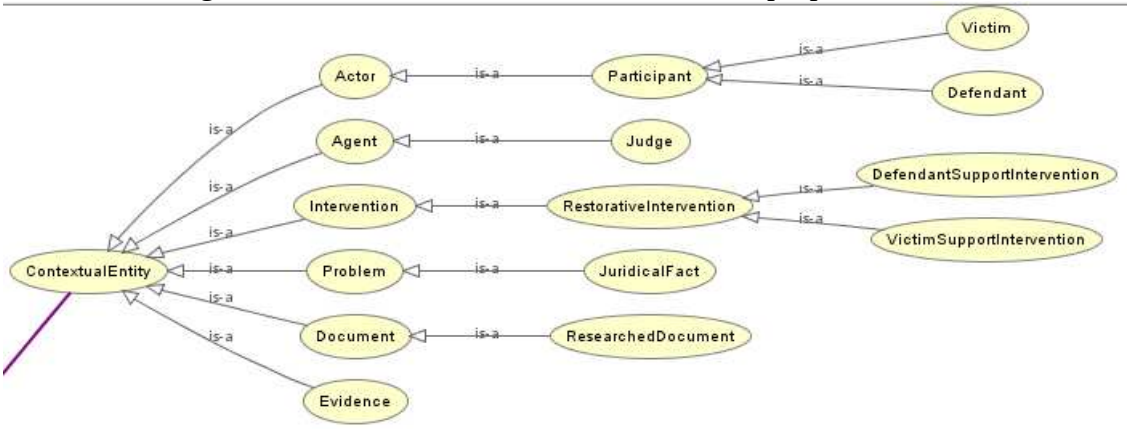


Figure 2. Subclasses of *ContextualEntity* for the crime prevention domain

Table 1. Descriptions of the main contextual properties - crime prevention domain

Subclass	**Contextual Properties**	**Description**
Judge	*hasExpertAffinity*	identifies a relation of expertise from the Judge profile on a given subject matter (e.g. crimes against children). It helps to identify mutual affinities among judges optimizing performance and reducing search results
Participant	*hasAbilities*	represents the defendant's or victim's skills, and is used to find mutual affinities with intervention programs (e.g. revenue). Reduce search results
Defendant	*hasPotentialRisk*	comes from juridical and psychosocial evaluations (profile). Behavior data, conduct, fact description and given sentences, especially for recurrent cases, are bases to characterize an offender's degree of risk.
JuridicalFact	*hasCircumstances*	describes the information about time and geographic aspects for the occurred fact. Information about number of people involved and their attitudes are desirable too. It is relevant and determinant to understand the juridical fact
Researched Document	*hasValidity*	indicates whether the document should be selected based on its quality and the methodological rigor associated with the question asked by the decision maker (judges, in this case)
	hasRelevance	indicates whether the set of results (outcomes) in the document, often presented in statistical form, is consistent and significant
Evidence	*hasApplicability*	indicates whether the evidence presented in the document is credible in the context of other knowledge, or whether it has practical utility in general
Restorative Intervention	*hasAdaptability*	indicates the degree of coherence in the application of evidence for the conducted behavior, needs and preferences of the defendant (or victim)
	hasSafety	denotes the percentage of safety that the judge have to apply the specific evidence to a particular participant (defendant or victim)
	hasExpectation	refers to the percentage of support expected from the use of evidence in relation to the participant (defendant or victim)

ConventionalEntity contains all the specific subclasses of the domain, which does not have a direct influence of context. They are *JuridicalResearch*, *Seek* and *JuridicalEvidenceProvider* as shown in Figure 3. For the *JuridicalResearch* subclass, the property *historic* should include general comments and the number of documents that were accepted and rejected, besides of the properties presented in Figure 1. In the

Seek subclass, properties about the researched document validity time must be considered. They are specific to instantiate the start and the end of the document validity found. The *JuridicalEvidenceProvider* subclass contains the follow properties: *name*, that mean the name of the Internet site accessed, and *homepage*, that detain the URL address.

Besides the enumerations presented in Figure 1, for the crime prevention domain, we add a subclass *BasicProgram* with the following values instantiated: *citizen*, *revenue*, *education*, *psychosocial*, *psychiatric* and *shelter*. This subclass serves to support the conventional intervention (no evidence-based) that exists in the courts. All subclass with enumerations are children of the *TypeClass* class.

Figure 3. Subclasses of *ConventionalEntity* for the crime prevention domain

The figures and table of this subsection were not fully developed in this article due to the limits of space.

4.2. Inference Rule

Some inference rules have been built and others are in development. In the following section addresses the applicability of some of them. For the sake of space we will not describe the properties of all classes in this paper. We will mention briefly the characteristics of some of them.

Intervention program rule for victims: When the participant is an offender, he has access to any program described in the basic ontology. However, when the participant is a victim, potential routes are: citizenship, psychosocial and psychiatric;

Rule to retrieve evidence-based solutions centered on the judge's specialty: from the expertise of a judge, logged in the system, the research solution space can be reduced based on: (i) his/her expertise, (ii) two specialties, or (iii) all experts;

Rule to find documents with evidence based in query keyword: the query terms must be confronted with the words found in researched documents; programs must order the most similar to facilitate the choice of the decision maker and a ranking as presented.

5. Application to the Crime Prevention Scenario

We present an example adapted from a real case involving an alternative penalty - a model for infractions that are of minor and moderately offensive potential (e.g., contravention, illegal weapon possession). It deals with a new modality, face-to-face restorative justice, in which a victim that suffered violence of an alcoholic offender receives support. A prototype, developed in *Java* language, interacts with a XML Database generated by Protégé ontology editor. The original data, approximately one

hundred cases, were extracted from conventional Court's Database. Figure 4 presents data for searching by evidence in the local database.

The High or Moderate Intervention Complexity is due to offender and victim need of treatment. The Judge's expertise in the new case is "drug crimes". We applied similarity cosine formula used in *Information Retrieval* for keyword similarity search between query and document with evidence.

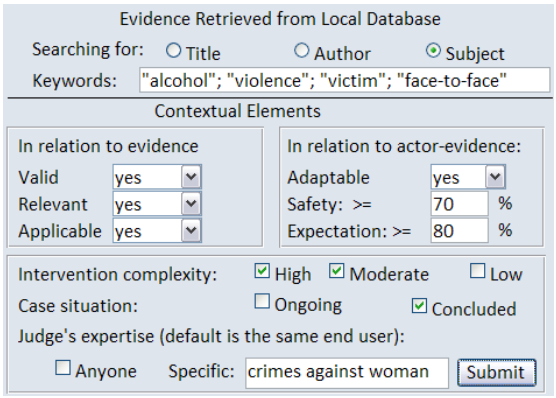


Figure 4. Data for searching evidence from local database

In the first retrieval, we do not use contextual elements and the results with several cases are present in Figure 5a. Using contextual information parameters as filter fewer cases were selected (see Figure 5b). This filtering was carried out as follows: (i) based on the desired expertise ("drug crimes" and "crimes against women") only the documents 1, 3, 4 and 5 were selected initially; (ii) document 4 was rejected by the safety indicator = 60.0 (so less than 70 % desired); and (iii) document 5 was not accepted by expectation indicator = 70.0 (so less than 80 % desired).

	title character varying(200)	keywords character varying(200)	study character	source charact	sa nu	ex nu	expertise character varyir	con cha	situ chai
1	Drunk and dangerous: a randomized controlled trial of alcohol	alcohol, brief interventions, violence, random	randomized	Spriger V	75	90	drug crimes	high	concl
2	Reducing violence through victim identification care and suppo	violence, crime victims rehabilitation, health p	narrative	World He	0.0	0.0	homicide		ongo
3	Assessing the effectiveness of interventions designed to supp	victims of crime, systematic review, violence,	systematic r	Campbell	80	85	crimes against wom	mod	concl
4	Change in behaviour of alcohol consumption: what is the moti	alcoholism, motivation, gastroenterology, out	case study	National .	60	50	drug crimes	mod	concl
5	Effects of Drug Substitution Programs on Offending among Dr	drug substitution, drug-addicts, alcohol depe	systematic r	Campbell	70	85	drug crimes	low	concl
6	Police crackdowns on illegal gun carrying: a systematic review	Campbell Collaboration, crackdowns, violence	systematic r	Spriger V	80	85	homicide	high	concl
7	Cognitive-Behavioural Interventions for Children Who Have Be	child sexual abuse, victim, cognitive-behaviou	systematic r	Campbell	85	80	crimes against child	mod	concl
8	School-Based Education Programmes for the Prevention of Ch	child sexual abuse, victim, school-based educ	systematic r	Campbell	70	80	crimes against child	low	concl

	title character varying(200)	keywords character varying(200)	study character	source charact	sa nu	ex nu	expertise character varyir	con cha	situ chai
1	Drunk and dangerous: a randomized controlled trial of alcohol	alcohol, brief interventions, violence, random	randomized	Spriger V	75	90	drug crimes	high	concl
2	Assessing the effectiveness of interventions designed to supp	victims of crime, systematic review, violence,	systematic r	Campbell	80	85	crimes against wom	mod	concl

Figure 5. Retrieved documents with evidence: a) without using context (upper), (b) using contextual element (lower)

The presented cases are not sufficient to give support to the solution (they do not treat face-a-face meeting). So, the judge should search for documents with evidences. The research began with the question containing the problem and actor (woman with a psychological problem who was assaulted), intervention (face-to-face sessions), comparison of interventions (face-to-face sessions and conventional processes) and outcome (beneficial effects). The sources Campbell Collaboration and Springer Verlag were chosen and their respective home-pages were obtained. Figure 6 show data for second search regarding documents published between 2005 and 2010.

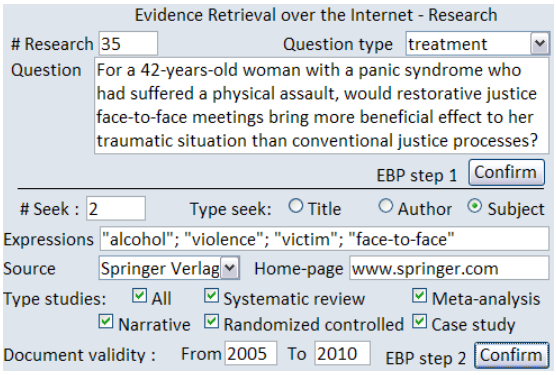


Figure 6. Data for searching evidence in Springer Verlag's database

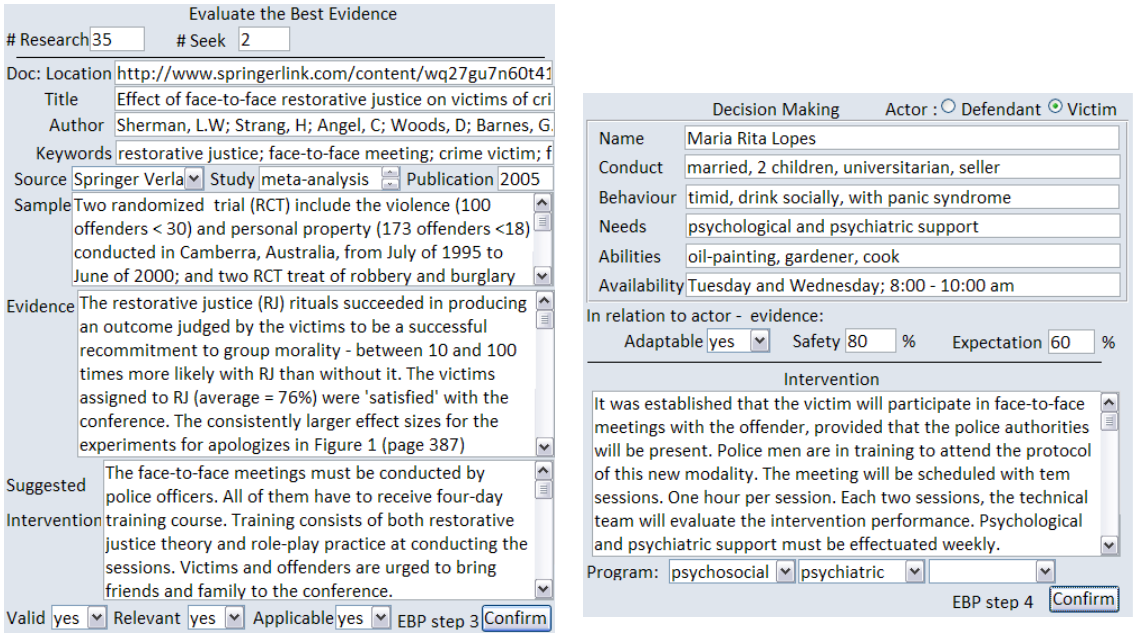


Figure 7. a) Evaluate the best evidence; e b) Decision-making g evidence in Springer Verlag's database

Figure 7a shows a meta-analysis study (taken from Springer Verlag) that was selected by the judge as presenting the best evidence on face-to-face meetings between victims and offenders. The study sample is derived from two randomized controlled trials: one conducted with offenders who committed crimes against private property involving violence in Canberra, Australia, and the other, crimes of burglary with victims in London, England. The sample context was analyzed and contrasted with the new problem. The evidence drawn from this showed objectively that 76% of victims were satisfied with the results obtained from the face-to-face meeting with offenders. This study led to a successful implementation of a training course for police officers, in which the concepts of restorative justice and practice sessions in face-to-face meeting between offenders and victims were applied. The document and the evidence were evaluated in terms of validity, relevance and applicability, and the information was extracted manually and recorded in a local database.

The decision making is presented in Figure 7b. Data of the victim were informed and they are compatible with the best evidence founded. The victim agrees to participate

in face-to-face meetings with the offender, provided that in previously established time and with the presence of authorities. Victim support programs, with respect to psychosocial and psychiatric treatment, must be offered in this particular intervention. The process concludes with documentation of the research performance made by judge.

This example shows that the presented application has potential to be leveraged to support a more appropriate evaluation of the ontology.

6. Related Works

In this section we present some related work on the themes *evidence, context, ontology,* and integration of this themes.

In Stolba et al. (2009) is showed how Data Warehouse facilitating Evidence-Based Medicine can be applied for reliable and secure processing of huge amounts of medical data. The authors present a data model for building a federated Data Warehouse considering adopted international standards for the exchange of healthcare data. Nakaya e Shimuzu (2006) present the Knowledge representation architecture based on Evidence based Logical Atomism (KELA) that consider the anatomic hierarchic structure from genome to human. Knowledge atoms of molecular and disease findings are modeled as entities and relationships - describes species, birthplace, and existing place in an ontological view.

Vieira et al. (2010) presents a domain-independent context meta-model, which guides context modelling in different applications. The meta-model offers integrated support for modeling structural and behavioral aspects involved in context management and usage. Contextual graph and UML were used. Sheng and Benatallah (2005) Introduce the ContextUML meta-model developed to support the modeling of context-aware Web Services. It separates modeling context (types, sources, etc.) from modeling context-awareness (objects and Mechanisms) becoming restrict to the Web Services category of Context-Sensitive Systems.

The related works above regard individually evidence or context. The combination of research evidence with context was not developed computationally. Besides, none of them has the perspectives of integration and extension for several domains, and none of these present a vision of combining ontological proposal.

7. Conclusions and Future Work

This article proposes the integration of context with evidence represented in a meta-model to facilitate the development of applications centered in EBP considering context for several domains. The class structure of the meta-model was the base for build domain ontology oriented to crime prevention. Contextual information related to the EBP of the criminal area were represented and instantiated. With a practical implementation we showed how contextual EBP can be used to support Judge's decision making and was verified that using contextual information makes the retrieve more effective.

Future researches encompass (i) the building of: task ontology for the criminal area; a high-level ontology for the areas that use EBP such as Medicine and Education; and a semi-automatic Evidence-Oriented Information Extractor (EOIE); and (ii) the

incorporation of the classical case structure (problem, solution and result) and Case-Based Reasoning technique for decision making support.

References

Brézillon, P. (2007). Context modeling: Task model and practice model. *CONTEXT-07*, LNAI 4635**.** pp. 122-135, Roskilde, Denmark.

Bunningen, A. (2004). "Context Aware Querying - Challenges for data management in ambient intelligence". Doctorate thesis, University of Twente.

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.

Dey, A.K. and Abowd, G.D. (2001). A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction (HCI) Journal*, v. 16, n. 2-4, pp. 97-166.

Dobrow, M.J., Goel, V. and Upshur, R.E.G. (2004). Evidence-based health policy: context and utilization. *Social Science & Medicine*, Jan, 58(1), 207-17.

Friedland, D. J., Go, A.S., Davoren, J.B., Shlipak, M.G., Bent, S.W., Subak, L.L. and Mendelson, T. (1998). Evidence-Based Medicine: A Framework for Clinical Practice. NY:McGraw-Hill.

Gomes, G.L.R. (2008). A Substituição da Prisão – Alternativas penais: legitimidade e adequação. Salvador: Editora *Podium.*

Nakaya, J. and Shimizu, T. (2006). Knowledge Architecture based on Evidence Based Logical Atomism for Translational Research. *International Journal of Computer Science and Network Security – IJCSNS*, February, v. 6, n. 2A, pp. 175-179.

Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W. and Haynes, R. B. (2001). Evidence-based medicine: how to practice and teach EBM. Elsevier Health Sciences.

Saliba, M.G. (2009). Justiça Restaurativa e Paradigma Punitivo. Curitiba: Juruá Editora.

Satterfield, J.M., Spring, B., Brownson, R.C., Mullen, E.J., Newhouse, R.P., Walker, B.B. and Whitlock, E.P. (2009). Toward a transdisciplinary model of evidence-based practice. *The Milbank quarterly*. Blackwell Publishing, June, v. 87, n. 2, pp. 368-390.

Sheng, Q. Z. and Benatallah, B. (2005). "ContextUML: A UML-Based Modeling Language for Model-Driven Development of Context-Aware Web Services". In Proc. of the International Conference on Mobile Business (ICMB05), pp. 206-212.

Stolba, N., Nguyen, T.M. and Tjoa, A. (2009) "Data Warehouse Faciliting Evidence-Based Medicine". In Nguyen, T.M. (Ed.). Complex DW and Knowledge Discovery for Advanced Retrieval Development, pp.174-195. Premier References Source.

Thomas, G. and Pring, R. (2004). Evidence-Based Practice in Education. Open University Press.

Vieira, V., Tedesco, P. and Salgado, A.C. (2010). Designing Context-Sensitive Systems: An Integrated Approach. *Expert Systems with Applications*, 38:2. pp.1119-1138.

Wang, X.H., Zhang, D.Q., Gu, T. and Pung, H.K. (2004). "Ontology based context modeling and reasoning using OWL". Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004.

Warren, R.K. (2007). "Evidence-Based Practice to Reduce Recidivism: Implications for State Judiciaries". http://works.bepress.com/roger_warren/1. April, 2009.

Using Multiple Views for Visual Exploration of Ontologies

Isabel Cristina Siqueira da Silva[1,2], Carla Maria Dal Sasso Freitas[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91501-970 – Porto Alegre – RS – Brasil

[2]Faculdade de Informática – Centro Universitário Ritter dos Reis (UniRitter)
CEP 90840-440 – Porto Alegre – RS – Brasil

{isabels@inf.ufrgs.br, carla@inf.ufrgs.br}

Abstract. *In this paper, we present a multiple views approach for exploring relationships between ontology classes and their instances. We employ thumbnails, 2D and 2.5D hyperbolic trees, which capture the hierarchical feature of parts of the ontology, while preserving the different categories of relationships between classes. In 2.5D visualization, the ontology is displayed as a tree on a plane, representing only the hierarchical relationships between concepts, and the user can explore other connections by creating projections of nodes (concepts) in another plane and linking them according to the relationships to be analysed. We report the comparison of our tool to Ontograf and propose a visualization method for the representation of instances of ontology classes.*

1. Introduction

Ontologies are used for sharing among people or software agents the common understanding of the information structure in a certain domain. As such, ontologies define concepts and ensure interoperability between systems. Gruber (1996) states that an ontology is a formal and explicit specification of a conceptualization. The conceptualization refers to the way people think, and the explicit specification relates concepts and relationships, which must be supplied in accordance with specific and well-defined terms.

However, due to the specificities of the concepts expressed in ontologies, the analysis of individual relationships is complex. Thus, the ontology visual representation and the quality of the provided interaction must be efficient. It is not simple to create a visualization that will display effectively all this information and, at the same time, allows the user to perform easily various operations on the ontology (Katifori et al., 2007). The challenge is to define the best way to represent relationships between categorized concepts, mainly because each concept can have a number of related relationships as well as attributes.

Visualization systems should consider two main issues: the mapping of information to a graphical representation in order to facilitate its interpretation by the users, and means to limit the amount of information that users receive, while keeping them "aware" of the total information space and reducing cognitive effort. When we analyze an image, we activate our perceptual mechanisms to identify patterns and perform segmentation of elements. The user must perceive the information presented on the display, and the understanding involves cognitive processes. An image can be

ambiguous or vacuous due to lack of relevant information or by excess of irrelevant information.

This work presents a visualization tool for exploring classes, instances and relationships in ontologies. In our previous works, we investigated ontology creation and visualization (Silva et al., 2009a, 2009b); performed requirements analyses and proposed a visualization tool based on interviews with experts who work with conceptual modeling and ontologies in two specific domains (Silva and Freitas, 2011a); and proposed a multiple views ontology visualization tool that aims at systematizing and transmitting knowledge more efficiently (Silva and Freitas, 2011b). In this paper we also report the comparison of our tool with Ontograf, which is also used for modeling and visualization of ontologies. The text is organized as follows. Section 2 briefly reviews related work. Section 3 summarizes our previous work and presents our new proposal. Finally, conclusions and future work are drawn in Section 4.

2. Related Works

Different alternatives for visualization and interaction with ontologies have been proposed (Katifori et al., 2007). In their work, Katifori el al. discuss different techniques that could be adapted for ontology representation, such as indented lists, trees and graphs, zooming, space subdivision (treemaps, information slices), focus+context and landscapes. Besides that, tools for ontology visualization and interaction are reviewed.

The OntoSphere tool (Bosca et al., 2005) uses two techniques - 3D visualization and focus+context – for providing overview and details according to user needs. Baehrecke et al. (2004) and Babaria (2004) proposed the use of treemaps to visualize data from GO (Gene Ontologies Consortium) database. In a treemap, color, size and grouping are used for facilitating user interaction and information extraction. Fluit et al. (2005) present the cluster map technique as a simple and intuitive method for visualizing complex ontologies.

Mostly, researchers use Protégé (Noy et al., 2000) for the creation and visualization of ontologies. Protégé's main visualization for the ontology hierarchy is a tree view (Class Browser). However, different visualization techniques have been proposed: Katifori et al. (2008) present a comparative study of four visualization techniques available in past versions of Protégé: Class Browser, Jambalaya (discontinued), TGVizTab (discontinued) and OntoViz (discontinued). The information retrieval features provided by these tools were evaluated.

The works by Samper et al. (2008) and Amaral (2008) address semantics aspects. Amaral (2008) proposes a semantics-based framework for visualizing descriptions of concepts in OWL. The framework aims at allowing users to obtain deep insights about the meaning of such descriptions, thereby preventing design errors or misconceptions. Icons and symbols are used in diagrams to characterize classes that represent concepts. One can combine information visualization techniques, as in the work by Schevers et al. (2006), where the user interacts with the ontology in the Protégé tool. Classes representing spatial information (like polygons, points, etc.) are presented in a second graphical interface that is used to mimic the functionality of a GIS (Geographic Information System).

Erdmann et al. (2008) presents the NeOn Toolkit, an open-source multi-platform environment, which provides comprehensive support for the ontology engineering life-

cycle. The toolkit is based on the Eclipse platform, and provides an extensive set of plug-ins (currently 45 plug-ins are available) covering a variety of ontology engineering activities. Catenazzi et al. (2009) present a study about tools for ontologies visualization and propose the OWLeasyViz tool. This tool combines textual and graphical representations for displaying class hierarchies, relationships and data properties. Interaction techniques such as zooming, filtering and search are available. Lanzenberger et al. (2010) discuss the visualization of ontology alignment as well as solutions for dealing with the inherent complexity of large ontologies. The presented techniques are also compared.

Recently, Kriglstein and Wallner (2011) presented Knoocks, a visualization tool focused on the interconnections between the ontology, its concepts and instances. This tool employs the overview + details approach, and was evaluated against another tools, although three of these tools are not available anymore in Protégé last versions (TGVizTab, OntoViz and Jambalaya). Also recently, Bach et al. (2011) proposed OntoTrix, a visualization technique designed to enable users to visualize large OWL ontology instance sets. The technique uses both node-link and adjacency matrix representations of graphs to visualize ontology data.

3. Multiple Views Tool

A multiple view system uses two or more distinct views to support the investigation of a single conceptual entity (Baldonado et al., 2000). Multiple views can help users understand complex relationships among different data sets. They are particularly helpful when coupling two or more views showing otherwise hidden relations.

As presented in section 2, many studies have addressed the importance of ontology visualization in creation, manipulation and inference processes. Different visualization methods have been proposed, but there are still many gaps to be filled in by efficient methods of visualization and interaction. The solution for these problems may be the simultaneous use of different techniques. In order to pursue an effective visualization tool, our study was divided in three steps:

- Interviews with four experts in creation and manipulation of ontologies to identify the requirements for an ontology visualization tool (Silva and Freitas, 2011a);
- Proposition and evaluation of a 2D and 2.5D hyperbolic tree visualization for exploring relationships between classes of ontologies (Silva and Freitas, 2011b);
- Proposition of a multiple views approach combining thumbnails and tree view visualization for exploring instances of classes in ontologies.

These three steps are described in details in the following sections.

3.1. Requirements Investigation

The study started with interviews with four people, all experts in the creation and manipulation of ontologies. The following questions were posed to the experts:

1. *Which aspects could be improved with visualization when an ontology is created?*

2. *Which information is searched more often after the ontology was created, and how this information could be displayed in order to make understanding more efficient?*
3. *Why (and when) is a visualization better than another?*

From the results of the interviews, we reached the following requirements for an ontology visualization tool:

- Provide overview of the ontology hierarchy, with the possibility of detailing some parts.
- Avoid presenting the different aspects of a specific ontology (classes, description, relationships, instances) together in a unique visualization.
- Optimize the results from ontology validation generated by inference processes.
- Explore the use of visual attributes such as color, transparency and shapes.
- Provide display filters based on different techniques of focus+context and/or overview+detail, zoom, pan and rotation of the image.
- Allow rapid and simple inclusion of visual elements in the visualization, as well as their removal.

These requirements were considered the starting point to propose our tool described below. There was also a last requirement – "Allow printing the entire ontology in paper sizes commonly used, such as A4", which was considered less important for the visualization design.

3.2. 2D and 2.5D Tree Visualization and Evaluation: Hierarchy and Relationships

We propose a visualization method that fits the requirements pointed out by users as well as the tasks listed by Katifori et al. (2007). In this step, we have chosen to focus on visualizing the hierarchy of the ontology and the relationships between concepts employing multiple views. For the hierarchy, we employ a 2D hyperbolic tree, a focus+context technique, which reduces the cognitive overload and the user disorientation that might happen during the interaction with the nodes, tree expansion and contraction, especially in ontologies with many concepts (Figure 1a).

However, besides the class hierarchy (relationship "is a"), users of ontologies need to analyze the other relationships in an integrated way. Thus, we use a second view to display a third dimension showing one or more relationships (object properties) selected by the user. To display them, we take the plane where the tree is displayed, and perform a 90° rotation around the X-axis (Figure 1b). The rotated plane, positioned in 3D as an XZ-plane, displays the hyperbolic tree, and selected relationships are represented as curved lines in space, connecting the related concepts, without interfering with the display of the hierarchical relationship.

Figure 1c shows the proposed 2.5D scheme applied to an ontology hierarchy/graph.

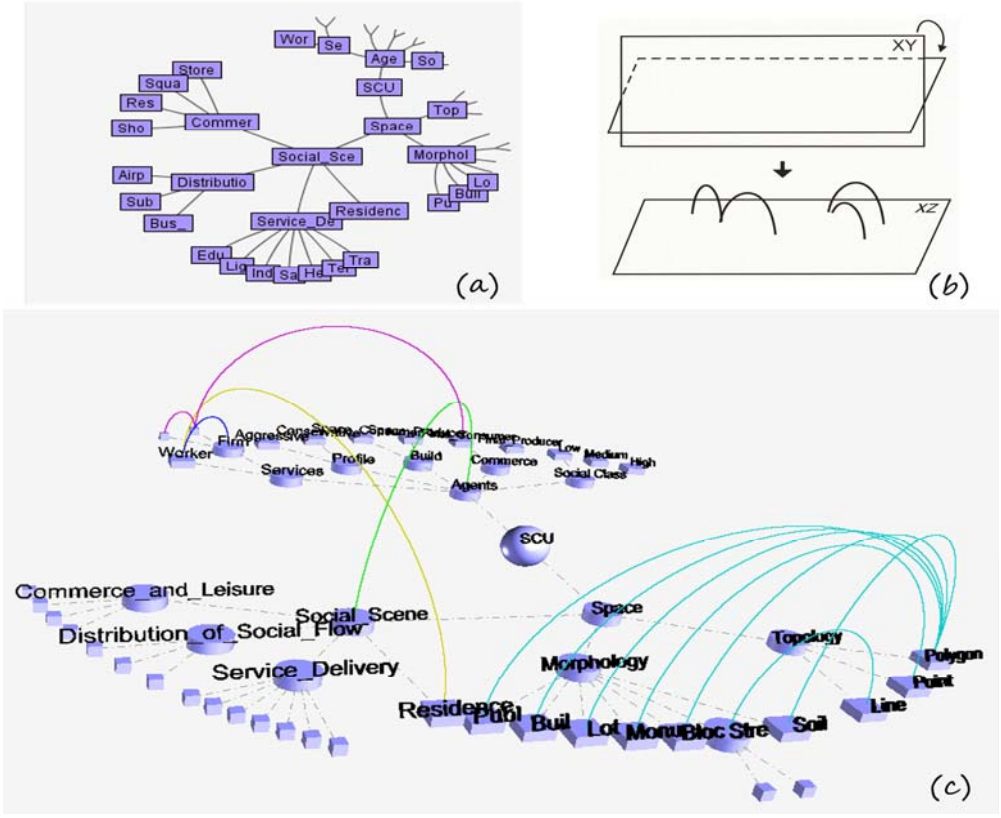


Figure 1. Ontology visualization. (a) 2D hierarchy visualization; (b) 2.5D visualization scheme; (c) Hierarchy and relationships in 2.5D

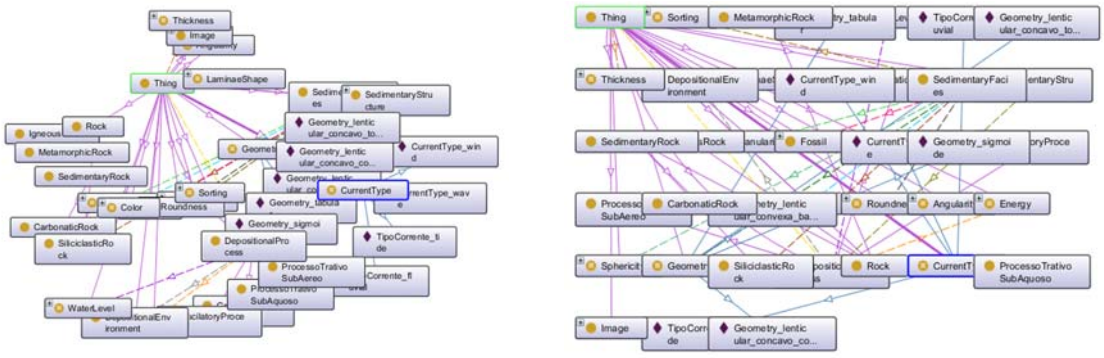
The main aspects of our technique regarding visualization and interaction are:

- Nodes are displayed with different geometric forms according to their type (root, subtree and leaf).
- Edges of hierarchy are displayed with solid lines and edges of relationships are displayed with dashed curves, the colors being related to the different relationships.
- In 2D hyperbolic tree view, the user can choose which nodes will be in focus on the image, hiding the other ones.
- Both 2D and 2.5D views can be displayed together, side by side, so the user remains "aware" of the ontology hierarchy and visualizes one or more relationships in a separate spatial dimension.
- The user can choose to display one or more relationships at the same time or hide them.
- In 3D space, the user can choose which levels of the tree view or hide, reducing the cognitive overload.
- In addition to rotations around the X-axis, rotations around the axes Y and Z, zoom and pan are also allowed, providing full 3D navigation.
- The background color can be changed.
- Tooltips are displayed over nodes and edges as additional information.

Such usability features aimed at reducing the cognitive effort of the user in analyzing the image and, at the same time, add functionality to the tool.

In order to evaluate our 2.5D visualization method, we have chosen to compare it with Ontograf (Falconer, 2010), a 2D tool for visualizing hierarchy and relationships of ontologies, which is available in the current version (4.1) of Protégé.

Ontograf presents seven visualization possibilities: alphabetical grid, radial and spring graphs, and four implementations of tree visualization: vertical, horizontal, directed vertical and directed horizontal. Figures 2a and 2b show two Ontograf visualization examples.



(a) (b)

Figure 2. Ontograf views (a) Radial; (b) Alphabetical grid

The four experts interviewed in the first phase of our study (as described in Section 3.1) were invited again to perform evaluations of our 2.5D visualization and Ontograf. Moreover, we invited two other experts in ontology specification to participate, so we had a sample of 6 people.

For the evaluations we used two ontologies as case study: a large ontology related to Stratigraphy concepts, and a smaller one, representing cities' urban performance. Before the participants started with the tasks, we shortly introduced them to the important functionalities of the tools, and they explored them in many ways using a training ontology. After the participants had finished their training, we started the evaluation process.

The tools were presented in different order for the participants. For each tool, they were asked to perform an analysis based on four questions that were defined in order to obtain the requirements listed in Section 3. The questions are listed below:

1. *Is the initial layout clear?*
2. *Is it possible to clearly separate the concepts' hierarchy from the other relationships between these concepts?*
3. *Does the possibility of rotating the ontology representation improve the analysis of relationships?*
4. *Do the pruning and expansion of the ontology levels enhance the understanding of hierarchical relationships?*

Three possibilities of answers were defined: Yes, Partially and No. Figure 3 summarizes the users' answers for these questions.

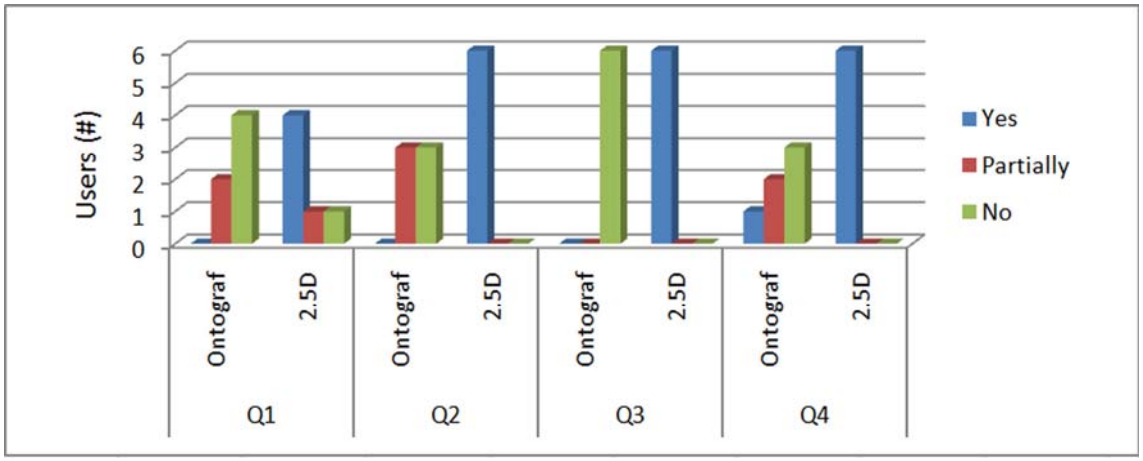


Figure 3. Evaluation results: Q1. Is the initial layout clear? Q2. Is it possible to clearly separate the concepts' hierarchy from the other relationships between these concepts? Q3. Does the possibility of rotating the ontology representation improve the analysis of relationships? Q4. Do the pruning and expansion of the ontology levels enhance the understanding of hierarchical relationships?

Regarding question (1), the majority of users (67%) responded that the initial 2.5D layout is clearer when compared with Ontograf. Among the reasons for that, users pointed out the large amount of information displayed at the same time (nodes overlap) in the image of Ontograf. This is a problem of scale versus amount of information, and causes user disorientation. In our 2.5D method, this problem is solved due to the nature of the hyperbolic tree.

In relation to question (2), users were divided (50%) between "Partially" and "No" answers for Ontograf, because nodes and edges overlap. Usually, users do not want to see relationships simultaneously, due to the cognitive overload that would arise. Thus, the possibility of analyzing the "is a" (hierarchy) and other relationships in different dimensions helps the user to understand the ontology. Another problem reported for Ontograf is that the user needs to change the positions of nodes in order to reveal the relationships occluded by them.

An important positive aspect noticed by users in both tools is the presence of tooltips when the mouse is over the nodes or relationships. The use of tooltip texts can help in the encoding of the displayed information, because they contain high loads of information, and are presented selectively as the user explores the visualization of the ontology.

Users also approved different colors for different types of relationships. Colors are mainly a resource for information categorization, and graphical elements like shapes and location of elements in the space help the user in mapping the concepts (Ware, 2008), and these features are present in our 2.5D method.

Regarding question (3), this functionality is not present in Ontograf, and the users considered it an important interaction mode. In our 2.5D method, rotations around the three axes (X, Y and Z) are possible, and complemented by zoom and pan. Thus, users have more freedom to interact with the visualization, and are able to reset to the

original layout at any moment. One of the users reported that when interacting with the 2.5D view, he did not feel claustrophobia, which is common in other tools, including Ontograf.

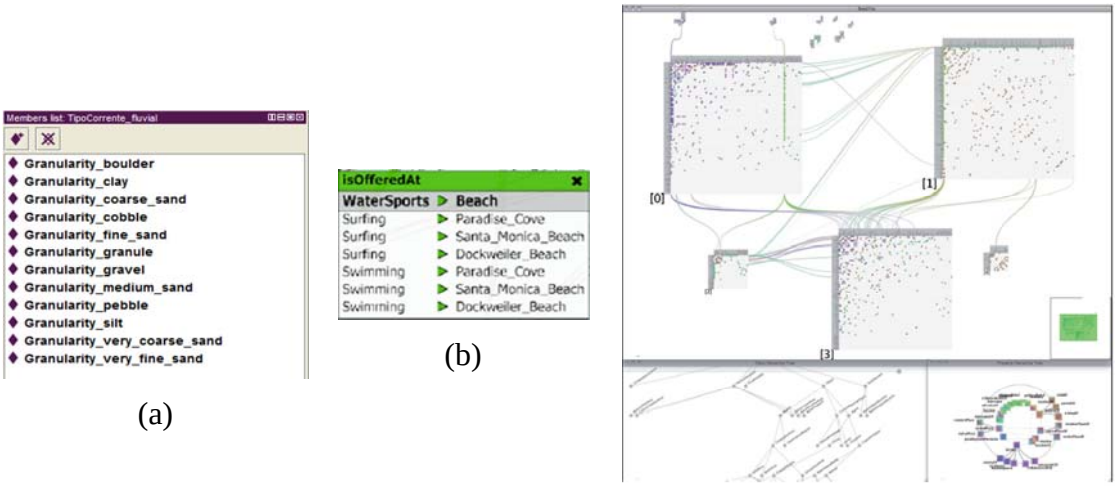
Finally, in relation to question (4), while 100% of users answered "Yes", for the 2.5D view, for Ontograf, most users (83%) answered "Partially" and "No". This result is due to the feature of Ontograf related to the repositioning of nodes when it is pruned or expanded, this fact causing disorientation on users. On the other hand, the 2.5D allows pruning and expansion in two ways: through the hyperbolic tree functionality of repositioning nodes, and through hiding/showing levels of the hierarchy.

These results indicate that the use of 2.5D visualization might be a solution to common problems presented by 2D and 3D ontology visualization tools, mainly cognitive overload and user disorientation.

3.3. Thumbnails and Treeview Visualization: Instances

Ontologies provide the explicit formalization and specification of the classes and their corresponding relationships (Gruber, 1996). Ontologies can have associated specific instances for the corresponding classes, which represent an essential part of any knowledge base, and are often orders of magnitude more numerous than the concept definitions. Thus, besides the class hierarchy and relationships, the visualization of instances of classes in ontologies is another important aspect.

The traditional way of displaying instances of classes in ontologies is through lists of items or indented lists (Figures 4a and 4b). We do not consider these the best modes of visualizing instances of classes because a lot of instances displayed together can generate cognitive overload. Bach et al. (2011) present OntoTrix (Figure 4c), which employed a hybrid visualization introduced as NodeTrix (Henry et al., 2007) in order to visualize the structure of ontologies. However, NodeTrix representations are less familiar to users than graphs and trees, making ontologies analysis more difficult and are likely to increase user's cognitive load.



(a) (b) (c)

Figure 4. Instances visualization. (a) Protégé (Noy et al., 2000); (b) Knoocks (Kriglstein and Wallner, 2011); (c) OntoTrix (Bach et al., 2011)

Based on the previous interviews with experts and in solutions adopted by other authors, we propose a visualization for showing instances of ontology classes, which employed thumbnails and a schematic treeview, as shown in Figure 5.

Figure 5. Proposed instances visualization

The classes of the ontology are exhibited as miniatures at the topmost part of the visualization (overview area), and the whole set can be navigated using shifting buttons as in common pictures displays. The detail area shows the selected class in treeview representation at the center of this view. This approach helps users to understand how the entire collection is organized, keeping both views visible for quick interaction. The classes can also be found by a search function of a typed keyword.

We adopted icons with a textual description because this representation gives a better comprehension than icons alone or text alone. The icons are related to two main categories of concepts of the urban ontology used in this example (agents and space) and are based on the work of Murray et al. (1994).

This schematic view is shown along two other views that show the 2D and 2.5D hyperbolic tree visualizations (Figure 6).

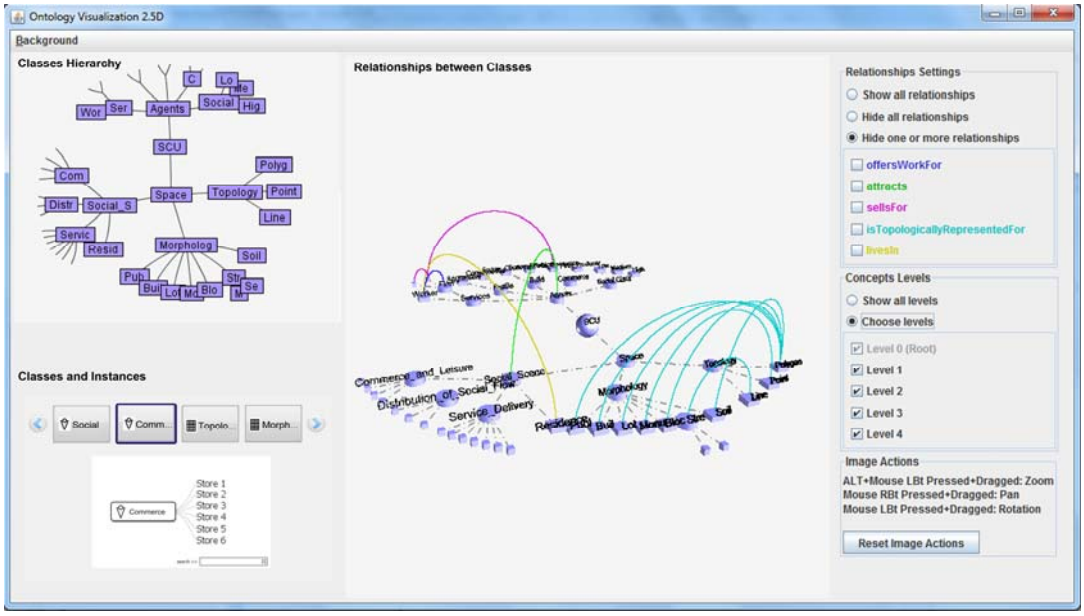


Figure 6. Multiple views in our tool

4. Conclusions

Ontologies tend to grow, incorporating new concepts and relationships, therefore increasing the visualization complexity. Static graphs, commonly used for ontology representation, are not the best alternative for such visualizations. Thus, we need efficient visualization and interaction methods tailored to ontologies. Information visualization techniques amplify cognition and reduce the exploration time of data sets, allowing the recognition of patterns and facilitating inferences about different concepts.

In this work, we have designed a visual and interactive method for exploring ontologies, aiming at improving the insight from such data. For this, we employed multiple views; a common and useful system that offer advantages like improved user performance, discovery of unforeseen relationships, and unification of the workspace.

We started this study with the definition of requirements for visualization and interaction with ontologies in order to support our design decisions for helping users to perform different operations on ontologies more easily and efficiently.

In our 2.5D visualization tool, we combine aspects of both 2D and 3D techniques. During its development we have taken into account the aspects pointed out by expert users. We evaluated the 2.5D visualization proposal by comparing it with the Ontograf tool, available in the version 4.1 of Protégé.

Besides the 2.5D visualization, we explore two other views: hierarchy classes and instances of classes. For the first, we use 2D hyperbolic tree, an intuitive focus+context technique that aims at representing very large trees. On the other hand, instances of classes are displayed as a hybrid view, exploring thumbnails and treeview methods. The main idea is to provide a visual representation that is intuitive and allows efficient analysis of the ontology concepts.

As future work, we intend to perform new evaluation experiments while investigating interactive visualization techniques for displaying data related to inferences.

Acknowledgements

We would like to thank the users that participated in the interviews and evaluation processes. Part of this work is financed by CNPq.

References

Amaral, F. Visualizing the semantics (not the syntax) of concept descriptions. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2008),* Vila Velha, ES, 2008.

Babaria, K. Using Treemaps to Visualize Gene Ontologies. Human Computer Interaction Lab and Institute for Systems Research, University of Maryland, College Park, MD USA, 2004.

Bach, B., Pietriga, E., Liccardi, I. OntoTrix: A Hybrid Visualization for Populated Ontologies. In: *20th Internacional Word Wide Web Conference*. Hyderabad, India. 2011.

Baehrecke, E. H., Dang, N., Babaria, K. Shneiderman, B. Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinformatics. 2004.

Baldonado, M., Woodruff, A. Kuchinsky, A. Guidelines for Using Multiple Views. In *Information Visualization*. Advanced Visual Interfaces, AVI, 2000. p. 110-119.

Bosca, A., Bomino, D., Pellegrino, P. OntoSphere: more than a 3D ontology visualization tool. In *Proceedings of SWAP, the 2nd Italian Semantic Web Workshop, Trento, Italy*, December 14-16, CEUR, Workshop Proceedings, ISSN 1613-0073, Vol-166, 2005.

Catenazzi, N., Sommaruga, L., Mazza, R. User-friendly ontology editing and visualization tools: the OWLeasyViz approach. In: *Proceedings of the 13th IEEE International Conference on Information Visualisation. Barcellona*, Spain. 14-17 July 2009. pp. 283-288. IEEE. ISBN: 978-0-7695-3733-7.

Erdmann, M., Peter, H., Holger, L, Studer, R. NeOn – Ontology Enggenering and Plug-in Development with the NeOn Toolkit. Url: http://www.neon-toolkit.org/images/tutorials/tutorial%20eswc08.pdf.

Falconer, S. OntoGraf. URL: http://protegewiki.stanford.edu/wiki/OntoGraf. Last access in 2010 october.

Fluit, C., Sabou, M., Harmelen, F. Ontology-based Information Visualisation: Towards Semantic Web Applications. In *International Symposium of Visualisation of the Semantic Web* (VSW'05). 2005.

Gruber, T. (1996). What is an ontology? [S.l.: s.n.], 1996. Url: http://www-ksl.stanford.edu/ kst/ what-is-an-ontology.html.

Henry, N., Fekete, J., McGuffin, M. Nodetrix: a hybrid visualization of social networks. IEEE TVCG, 13(6):1302–1309,2007.

Katifori, A.; Halatsis, C.; Lepouras, G.; Vassilakis, C.; Giannopoulou, E. Ontology visualization methods - a survey. ACM Comput. Surv. 39, 4 (Nov. 2007), 10.

Katifori A, Torou E, Vassilakis C, Lepouras G, Halatsis C: Selected results of a comparative study of four ontology visualization methods for information retrieval tasks. In: *Research Challenges in Information Science, 2008 RCIS 2008 Second International Conference on: 2008;* 2008: 133-140.

Kriglstein, S. Wallner, G. Development Process and Evaluation of the Ontology Visualization Tool Knoocks - A case study. In: *International Conference on Information Visualization Theory and Applications IVAPP, 2011*, Vilamoura-Algarve. Proceedings of the International Conference on Imaging Theory and Applications and International Conference on Information Visualization Theory and Applications. Portugal: SciTePress Science and Technology Publications, 2011. p. 187-197.

Lanzenberger, M., Sampson, J., Rester, M. Visualization in Ontology Tools. Ontology Visualization: Tools and Techniques for Visual Representation of Semi-Structured Meta-Data. Journal of Universal Computer Science, vol. 16, no. 7 (2010), 1036-1054.

Murray, N., Paton, N., Goble, C., Bryce, J.. Kaleidoquery: a flow-based visual language and its evaluation. Journal of Visual Languages and Computing,11:151-189, 2000.

Noy, N., Fergerson, R., Musen, M. The knowledge model of Protege-2000: Combining interoperability and flexibility. In *Proceedings of 2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000*), Juanles-Pins, France, 2000.

Samper, J., Tomás, V., Carrillo, E., Nascimento, R. Visualization of ontologies to specify semantic descriptions of services. IEEE Transactions on Knowledge and Data Engineering. 20(1): p. 130-134. 2008.

Schevers, H.A.J., Trinidad, G.; Drogemuller, R.M. Towards Integrated Assessments for Urban Development. Journal of Information Technology in Construction (ITcon), Vol. 11, Special Issue Decision Support Systems for Infrastructure Management, pg. 225-236. Url: http://www.itcon.org/2006/17.

Silva, I. C. S., Netto, V.M., Freitas, C. M. D. S. Novos caminhos para simulação urbana: integrando métodos de visualização de Informações e modelagem de agentes e redes espaciais In: *XIII Congreso Iberoamericano de Gráfica Digital - Sigradi*, Sao Paulo - SP. 2009 (a)

Silva, I. C. S., Freitas, C. M. D. S., Netto, V.M. Ontologia para Sistemas Configuracionais Urbanos In: *II Seminário em Ontologia no Brasil, (Ontobras)*, Rio de Janeiro. 2009. (b)

Silva, I. C. S., Freitas, C. M. D. S. Requirements for Interactive Ontology Visualization - Using Hypertree+2.5D Visualization for Exploring Relationships between Concepts. In: *International Conference on Information Visualization Theory and Applications IVAPP, 2011*, Vilamoura-Algarve. Proceedings of the International Conference on Imaging Theory and Applications and International Conference on Information Visualization Theory and Applications. Portugal: SciTePress Science and Technology Publications, 2011. p. 242-248. (a)

Silva, I. C. S., Freitas, C. M. D. S. Using Visualization for Exploring Relationships between Concepts in Ontologies. In: 15th International Conference on Information Visualisation, IV 2011, 2011, London, UK. Information Visualization: Visualization, BioMedical Visualization, Visualization on Built and Rural Environments & Geometric Modelling and Imaging, 2011. p. 317-322. (b)

Ware, Colin. Visual Thinking for Design. Morgan Kaufmann, Burlington, MA, 2008.

Ontology to Classify Learning Material in Software Engineering Knowledge Domain

Joselaine Valaski, Andreia Malucelli, Sheila Reinehr, Ricardo Santos

Programa de Pós-Graduação em Informática (PPGIa)
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba – PR – Brasil

{jvalaski, malu}@ppgia.pucpr.br, sheila.reinehr@pucpr.br,
ricardo.c.r.santos@gmail.com

Abstract. *This paper proposes an ontology to automatic classification of learning materials to the Software Engineering knowledge domain. The Software Engineering Body of Knowledge (SWEBOK) was used to define the hierarchical structure of the knowledge area. The Rational Unified Process (RUP) was used to add the axioms to represent the relationships between concepts and to enable the reasoning to SWEBOK knowledge areas. Two testing scenarios were designed and experiments were performed. The results show that the ontology is able to classify and locate learning materials from the Software Engineering area, according to the desired area, role, artifact or task.*

1. Introduction

The development of new web-based technologies has increased the number of learning environments, from simple learning resources repositories to more complex learning environments. In these environments, learner can access information, communicate among themselves and learn in a self-learning method [Ruiz et al. 2008].

This self-learning process can happen through many didactic materials, such as digital books, slideshows, audio or video recordings, etc. These materials allow knowledge sharing within a common interest domain and are available to anyone, anytime, anywhere. This can facilitate the learning of subjects that require highly trained professionals, who need to be up-to-date with the state of the art of technology. Software Engineering can be named as one of such subjects.

However, the self-learning environment can present challenges that hinder the real knowledge acquisition. The difficulty to search the learning materials according to the learning theme is one of these challenges. This search can be more difficult to learners due to the range of knowledge themes [Yu 2010], making the identification of desired learning materials a challenge [Fischer 2001].

The process of classifying learning materials according to their knowledge area can be an alternative to facilitate their retrieval. However, these classification mechanisms must use a common language that would allow knowledge sharing to occur effectively [Davenport and Prusak 1998].

Most knowledge areas have terminology problems in the use of consensual terms, as an example, the Software Engineering area. It is common that different development teams use diverse terms for the same concepts. Even though many software engineers work with Software Engineering, some professionals claim to never have studied the subject [Wongthongtham 2006]. Thus, it is likely that professionals find some difficulty to search adequate learning materials due to lack of a common terminology.

In this context, ontologies play an important role because they can be applied to provide a common shared understanding of an information structure among individuals or organizations, as

well as be used to enable the knowledge domain reuse and make explicit assumptions of a domain [Noy and McGuinness 2010].

Ontologies can describe a hierarchy of concepts related by subsumption relationships, in this case, a taxonomy-driven concept; or a structure, where the axioms are added in order to express relationships between concepts and to restrict their intentional interpretations [Guarino 1998]. Through ontologies, hierarchical structures of themes related to the learning materials can be defined using a common vocabulary to the knowledge area. Furthermore, it is possible to add reasoning to this structure in order to help the automatic classification of learning materials within the defined hierarchy. The automated classification is relevant when people do not hold enough knowledge to identify the theme related to the learning materials due to lack of common vocabulary of the knowledge area. Software engineers can be mentioned as an example.

In this context, this paper aims to propose an ontology to automatic classification of learning materials related to the Software Engineering knowledge area. The ontology aims to facilitate the search for learning materials within the given domain. The Software Engineering Body of Knowledge (SWEBOK) [Abran and Moore 2004] was used to define the hierarchical structures of knowledge. The SWEBOK is intended to reach broad consensus on the area of Software Engineering [Sicilia 2005]. The Rational Unified Process (RUP) was used to add axioms to represent the relationships between concepts and enable the reasoning to the SWEBOK knowledge area.

The remainder sections of this paper are organized as follows: Section 2 presents the related work; Section 3 describes in details the proposed ontology; in Section 4 some experiments are discussed; Section 5 concludes the paper.

2. Related Works

There are several papers proposing ontologies for the Software Engineering area. This section presents these researches and their approaches.

Mendes and Abran (2005) present a prototype of an ontology to represent the domain of Software Engineering, based on the SWEBOK guide. A literal extraction from the guide results in approximately 4,000 concepts. In this approach, there is no intention to establish a hierarchical structure of the Software Engineering knowledge area. Sicilia et al. [2005] also proposes a SWEBOK based ontology with a descriptive part in order to identify artifacts and activities and a prescriptive part, with approaches and concrete activities'rules for "commonly accepted" practical activities. Hilera et al. (2005) propose an ontology called OntoGLOSE based on the Software Engineering Terminology Glossary, published by IEEE. OntoGLOSE includes about 1,500 concepts, corresponding to 1,300 glossary terms with their different meanings.

More specific approaches are established on the Software Engineering domain as well. The Win-Win approach represents a model created to manage the necessary collaboration and negotiation by the people involved in the software lifecycle stage [Bose 1995]. ONTODM represents the knowledge of requisite specification techniques of a multi-agent systems family in an application domain. It is being used as a CASE tool to help to elicit and specify the domain models. [Girardi and Faria 2003]. Sánchez et al. [2005] propose an ontology to represent the different meanings of the term model, incorporating the different concepts related to the terms. Cyc [2011] presents a UML subOntology integrated in the OpenCyc ontology containing about 100 concepts, 50 relationships and 30 instances, including UMLModel Element, UMLClassifier, UMLClass and UMLStateMachine, according to SWEBOK's Software Projects Notations subarea, from the Software Project area. The XCM ontology provides a pattern to a component definition that appears in different component models and standardizes these differences [Tansalarak and Claypool, 2004]. Deridder [2002] presents a general ontology on concepts related to software maintenance. An ontology organized in five subontologies to represent the knowledge related with software systems,

the necessary skills to software maintainers, with maintenance process activities, organizational maintenance topics and tasks that constitute any application domain is proposed by Dias et al. [2003]. Ruiz et al. [2004] propose an ontology composed by four subontologies: products, activities, organization processes and agents. Vizcaino et al. [2005] propose an ontology composed by the ontologies proposed by Deridder [2002], Dias et al. [2003] and Ruiz et al. [2004]. The propose of Deridder [2002], Dias et al. [2003], Ruiz et al. [2004] and Vizcaino et al. [2005] are based on an initial software maintenance ontology proposed by Kitchenham et al. [1999]. Boehm and In [1996] propose an ontology with concepts related to software quality attributes and information about the software architectures influences and development processes on these attributes Other ontology related to software process concepts is proposed by Falbo et al. [2002]. An ontology with the software measurement terminology, associated with fundamental concepts is proposed by Garcia et al. [2005]. Tautz and Greese [1998] present an ontology of the GQM (Goal Question Metric) paradigm, and an ontology with concepts related to software process, including Life Cycle Models concepts, Software Processes, Activities, Procedures, Tasks, Roles or Artifacts is presented by Falbo et al. [1998]. The SPOnt, an ontology that reused concepts from other ontologies related to decision support systems, establishing relationships, is proposed by Larburu et al. [2003]. González-Pérez and Henderson-Sellers [2006] present an ontology for software development methodology that include a metamodel and an architecture divided into three domains. Lin et al. [2003] propose an ontology for the IEEE 12207 and the CMMI Standards that can be applied in an organization in order to inspect and enhance the software processes maturity. An ontology particularly focused on the Software Engineering area was developed by Wongthongtham et al. (2007), the first Software Engineering oriented ontology, based on the SWEBOK's areas of knowledge. This ontology presents only a hierarchical structure; it does not use axioms to define the concepts related to the knowledge areas.

There are several proposals for ontologies in the Software Engineering area, however, there is not an ontology to classify materials according to the Software Engineering knowledge area. The next section discusses the proposal of an ontology to help solving this problem.

3. Proposed Ontology

This section presents an ontology composed by SWEBOK and RUP concepts to classify learning materials in the Software Engineering knowledge area. The ontology was developed with the ontology editor Protégé [Stanford 2011].

To define the knowledge's hierarchical structure related to Software Engineering, the SWEBOK´s definition knowledge area was used. The SWEBOK is a guide created under the patronage of the Institute of Electrical and Electronics Engineers (IEEE) with the objective of serving as reference to Software Engineering related subjects [Abran and Moore 2004]. This guide presents a hierarchical classification of the Software Engineering topics, where the higher level is the knowledge areas.

However, the definition of a hierarchical structure is not enough to allow the automatic classification of learning materials according to the defined structure. The SWEBOK does not present an approach to the definition of their knowledge areas using relationships among the concepts or explicit properties. For this reason, RUP was also used. RUP presents well-defined relationships among the main concepts, which are: Discipline, Artifact, Role and Task. Although RUP is a software development process, hence, not exactly focused on knowledge areas, the concept of disciplines can be related between some SWEBOK knowledge areas, as shown in Table 1. In this proposal, only the areas with total correspondence were mapped.

In the following subsections the details of the proposed ontology for RUP and the integration of this ontology with the ontology for the classification of learning materials according to the SWEBOK knowledge areas are presented.

Table 1 – Relationship between the SWEBOK areas and RUP disciplines

SWEBOK Area	RUP Discipline
Software Engineering Management	Project Management
Software Engineering Process	
Software Engineering Tools and Methods	
Software Configuration Management	Configuration and Change Management
Software Construction	Implementation
Software Design	Analisys and Design
Software Maintenance	
Software Quality	
Software Requirements	Business Modeling Requirements
Software Testing	Test
	Deployment
	Environment

3.1 OntoRUP: RUP representation ontology

OntoRUP was developed according to the Artifact, Role and Task concepts and their relationships with the Discipline concept. Through these four concepts and their relationships, classes and their properties were created. Table 2 presents the created classes and properties.

Table 2 – Classes and properties from OntoRUP

Domain Class	Range Class	Property	Special Property (inverse)
Artifact Task	Discipline	hasDomain	isDomainOf
Discipline	Artifact Task	isDomainOf	hasDomain
Role	Artifact	modify	isModified
Artifact	Role	isModified	modify
Task	Role	hasPerformer	isPerformerOf
Role	Task	isPerformerOf	hasPerformer

The general proposed hierarchy is presented in Figure 1. The RupElements class was created in order to group the derivative concept classes: Discipline, Artifact, Role and Task concepts.

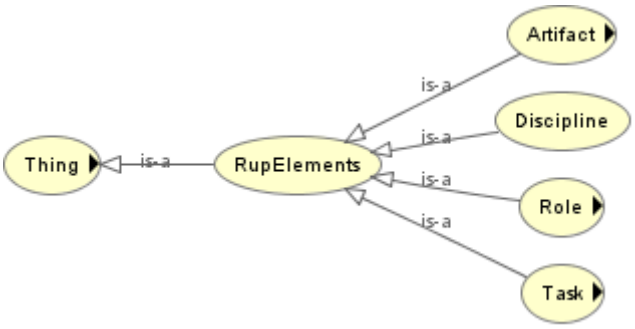


Figure 1 – OntoRUP general hierarchy

The Discipline class was created to represent the nine disciplines that compose the RUP model. Through this class the other relationships are established and then the integration is done with the SWEBOK´s knowledge areas.

The Artifact class was created to represent the software artifacts that are used within the RUP process. The Artifact class is directly related to the Discipline class through the hasDomain property. According to this relationship, subclasses were created, that identify the artifacts related to each of

the nine disciplines proposed in the RUP model. Figure 2 presents an example of the hasDomain property.

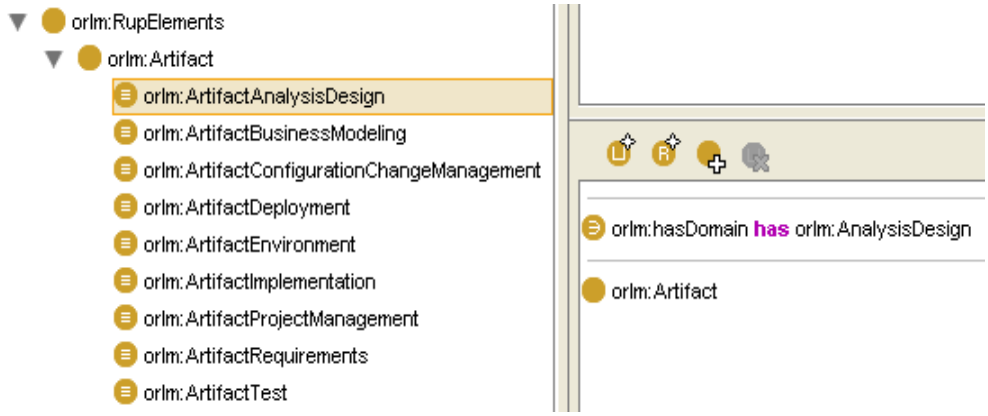


Figure 2 – hasDomain property

The Role class was created to represent the corresponding subclasses to the six groups of roles within the RUP, namely: Analysts, Developers, General Roles, Manager, Production Support and Testers. Furthermore, within the Role class, corresponding subclasses of the roles related to each of the nine disciplines were also created as shown in Figure 3. To establish the relationship between the Role and Discipline classes, it was used the property "modify" that relates the Role class to the Artifact's subclasses. As the subclasses of Artifact are already related to the Discipline class, the relationship between the Role and Discipline classes is also completed.

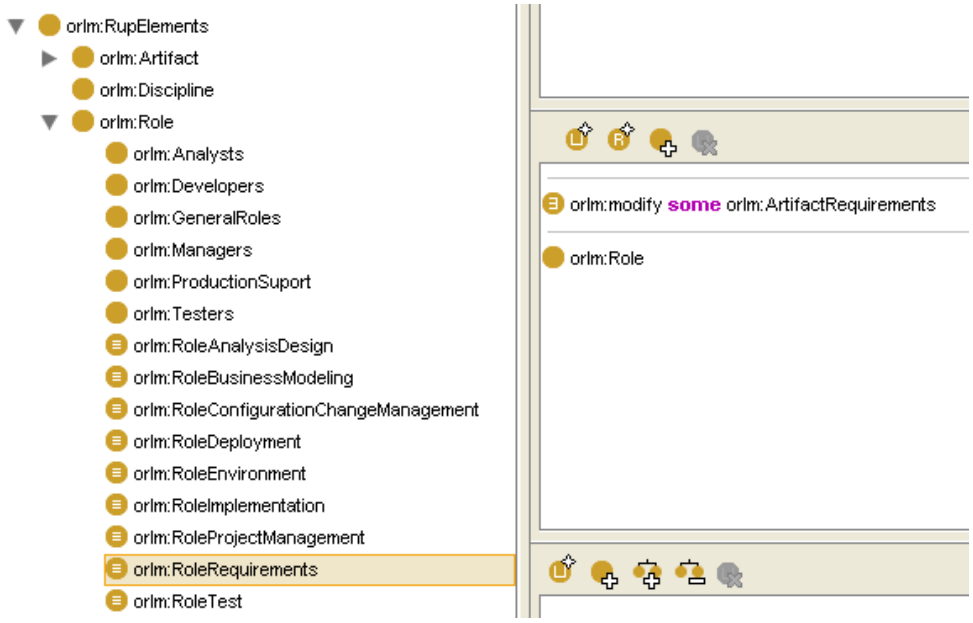


Figure 3 – Role's subclasses

The Task class was created to represent the tasks of the RUP model. The Task class has direct relationship with the Discipline class through the hasDomain property. Based on this relationship have been created subclasses to represent the tasks corresponding to each of the nine RUP disciplines specified in the model.

3.2 Software Engineering Learning Materials Ontology

Once established the ontology structure for representation of RUP elements, it was defined the necessary elements to enable the classification of learning materials within the Software Engineering domain. The LearningMaterial class was created to represent the learning materials, and its subclasses were created based on the ten SWEBOK's areas, as shown in Figure 4.

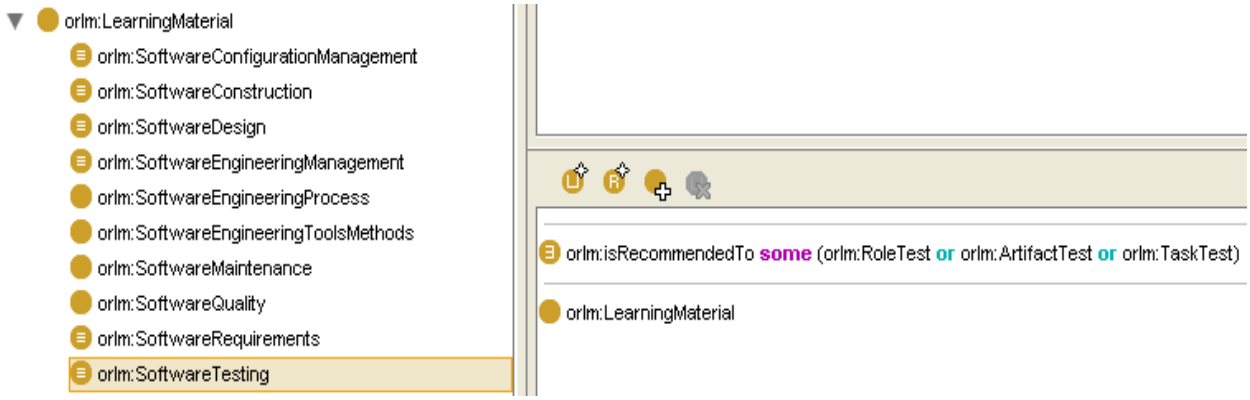


Figure 4 – Learning Materials according to the SWEBOK

In order to define the ten areas of the SWEBOK using explicit and formal properties, the defined concepts of RUP ontology was used. It is possible to identify the related discipline through any of the Artifact, Role and Task concepts, and through mapping it is possible to know the SWEBOK´s knowledge area. Because of that, the isRecommendedTo property was created, as shown in Table 3, in order to be able to recommend a learning material related to any of the three concepts presented in RUP. Thus, when adding a learning material it is possible: to recommend the material for the use of a specific artifact, such as a Business Case; the execution of a specific task, such as Architectural Analysis; or the execution of a specific role, such as System Analyst.

Table 3 – isRecommendedTo property

Domain Class	Range Class	Property	Special Property (inverse)
LearningMaterial	Artifact Task Role	isRecommendedTo	hasRecommendation
Artifact Task Role	LearningMaterial	hasRecommendation	isRecommendedTo

Through the related recommendation it is possible to classify the material according to the SWEBOK's knowledge areas. For instance, a learning material will be classified as belonging to the Test knowledge area, if it has the isRecommendedTo property related to, at least, one instance of the Artifact, Role or Task classes, linked to the Test discipline.

These possibilities of recommendations can help to obtain a more accurate classification of the learning material, especially when there is no formal knowledge regarding to which knowledge area the material belongs to.

4. Results

The ontology was proposed to be applied in a self-learning environment where people share their knowledge related to the Software Engineering area by adding learning materials. The proposed ontology will help in the classification of learning materials, mainly because software engineers may not use a common vocabulary or may not have enough knowledge to classify correctly the material

within the appropriate domain. Furthermore, the ontology will facilitate the recommendation of these learning materials.

Two scenarios were designed to verify the proposal's viability. The scenario 1 was used to test the classification of learning materials and scenario 2 was used to test the recommendation of these materials. The simulations were created using the Protégé tool.

- Scenario 1 – Learning Materials Classification

Instances of learning materials were added using the Protégé tool, as shown in Figure 5. Also recommendations were made through the isRecommendedTo property. Each recommendation was associated with instances of Artifact, Role or Task classes.

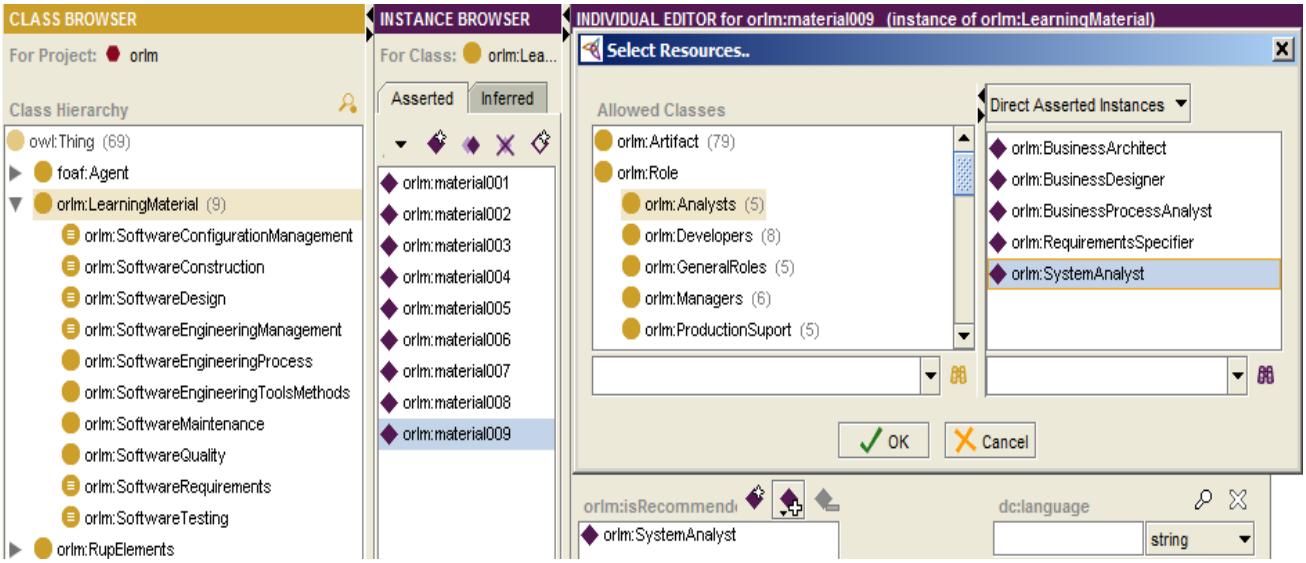


Figure 5 – Learning materials instances included using Protégé

The values assigned to the isRecommendedTo class for each one of the learning materials are shown in Table 4.

Table 4 – Values assigned to the isRecommendedTo property

Id. Material	Recommendation for Artifact	Recommendation for Role	Recommendation for Task
material001	Analisys Model Use Case Model	System Analyst	
material002		Requirements Specifier	
material003			Create Baseline
material004		System Administrator	
material005	Test Plan		
material006			Architectural Analisys
material007		Software Architect	
material008	Business Case	System Analyst	
material009		System Analyst	

The Pellet reasoned, version 1.5.2, was used to classify the learning materials. As shown in Figure 6, it is possible to verify that the ontology correctly classified the learning materials according to the defined concepts.

However, it is important to provide mechanisms to help software engineer to make their recommendations in order to avoid inconsistencies. For instance, the material identified as "material008", was recommended to be used in the Business Case artifact. In this case, it should not be possible to recommend it for the System Analyst role, as this role has no relationship with this artifact. As a result, the material was classified in three knowledge areas, one of them due to artifact

recommendation, and the other two due to recommendation by role. The ontology proposed can be used to help filter consistent recommendations among Artifact, Role and Task classes.

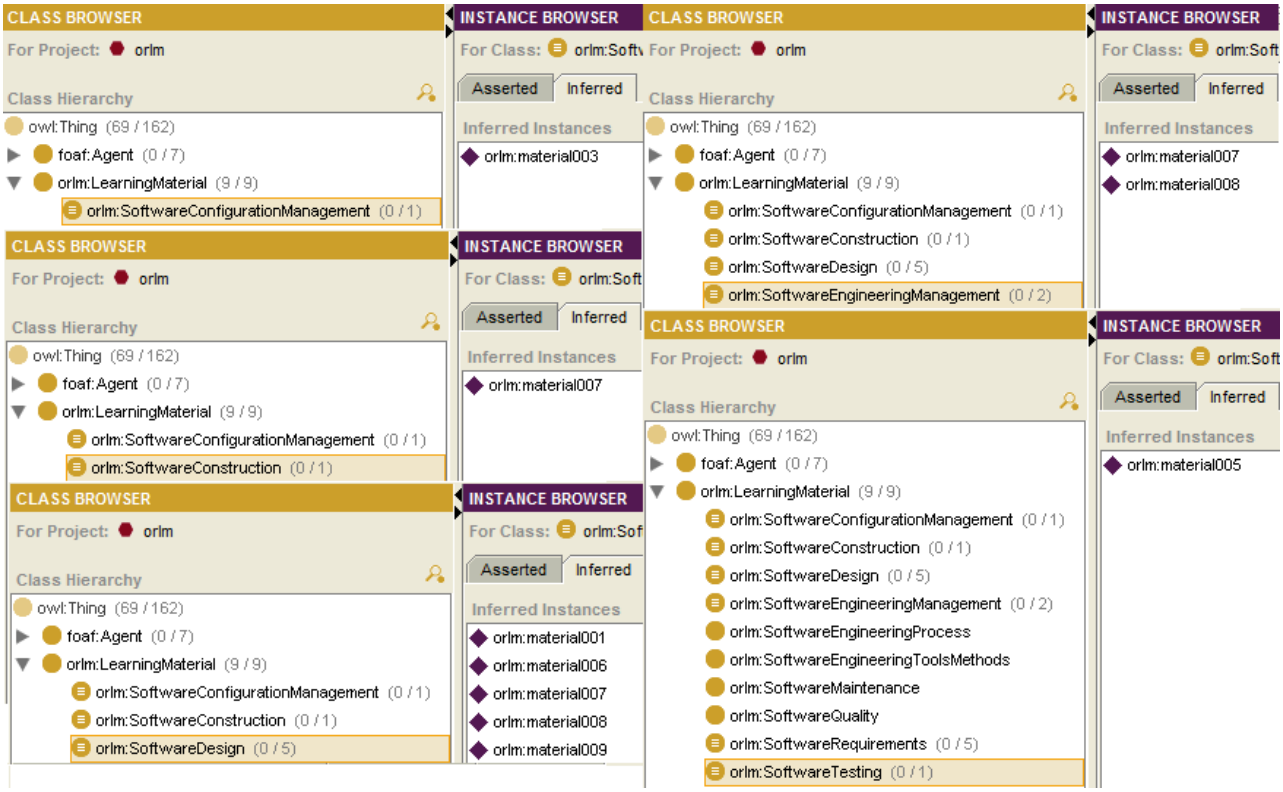


Figure 6 – Learning Materials classification using Pellet

- Scenario 2 – Learning Materials Recommendation

Scenario 2 was designed to present the possible recommendations of the learning materials once these materials will be available in a learning environment. According to the simulation described in scenario 1, after the learning materials were classified using the inference mechanisms, it is possible to search for these materials through the knowledge areas defined in SWEBOK. For instance, it is possible to retrieve all the learning materials related to the Software Requirement area. However, besides retrieving the materials by Software Engineering knowledge area, the ontology also allows to find all the materials according to recommendations, by Artifact, Role or Task.

SPARQL was used to simulate a preview of these possibilities. The SPARQL is a language to retrieve data from Web Ontology Language (OWL) files. Figure 7 presents a SPARQL query in order to retrieve learning materials recommended by Roles. In this case, the learning materials are retrieved through the Roles view; however, the queries can be executed by Artifacts and Tasks as well.

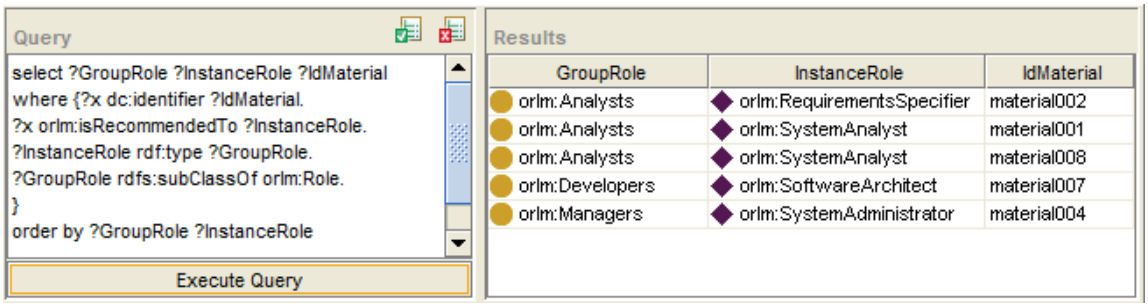


Figure 7 – Query using SPARQL

It is important to point out that new recommendations may be added to the learning materials according to their use. For example, a learning material that was added with the System Analyst role may also be recommended to the Elicit Stakeholder Requests task. So, the level of details for the recommendation is enhanced and the retrieval of material becomes more precise.

5. Conclusion

This paper presented an ontology to automatically classify learning materials related to the Software Engineering knowledge area, aiming to facilitate the search for these materials.

The ontology was defined using the main structure of ten SWEBOK knowledge areas and the concepts and relationships among Artifact, Task and Role elements from RUP model. RUP was used to define SWEBOK knowledge areas through axioms to enable the automatic classification of learning materials according to recommendations.

Some experiments were performed and it was possible to conclude that the ontology classifications were correctly, according to the Software Engineering knowledge areas. Furthermore, the ontology provides views of the learning materials under three aspects, recommendations by artifacts, tasks and role. This diversity can be another facilitator for retrieving the desired material.

The proposed ontology will be integrated to a self-learning environment, and experiments with Software Engineering students and professionals will be performed in order to evaluate the proposal.

References

Abran, A. and Moore, J. W. (2004). "SWEBOK - Guide to the Software Engineering Body of Knowledge". IEEE CS Professional Practices Committee.

Boehm, B. and In, H. (1996). "Identifying Quality Requirements Conflicts". IEEE Software, pp. 25–35.

Bose, P. (1995). "Conceptual design model based requirements analysis in the Win-Win framework for concurrrent requirements engineering". In: IEEE Workshop on Software Specification and Design (IWSSD).

Clemente, J., Ramírez, J. and Antonio, A. (2010). "A proposal for student modeling based on ontologies and diagnosis rules". Expert Systems with Applications, pp. 8066-8078.

Cyc (2011). Cyc: OpenCyc.org: Formalized Common Knowledge. Cycorp, USA. http://www.opencyc.org, April.

Davenport, T.H. and Prusak, L. (1998). "Working Knowledge: How Organizations Manage What They Know". Harvard Business School Press.

Deridder, D. (2002). "A Concept-Oriented Approach to Support Software Maintenance and Reuse Activities". In: 5th Joint Conference on Knowledge-Based Software Engineering (JCKBSE), Maribor, Slovenia.

Dias, M.G., Anquetil, N., and Oliveira, K.M. (2003). "Organizing the Knowledge Used in Software Maintenance". In: Journal of Universal Computer Science, pp. 641–658.

Falbo, R., Menezes, C. and Rocha, A. (1998). "Using Ontologies to Improve Knowledge Integration in Software Engineering Environments". In: 4th International Conference on Information Systems Analysis and Synthesis(ISAS), Orlando, USA.

Falbo, R.A., Guizzardi, G., Duarte, K.C. (2002).”An Ontological Approach to Domain Engineering”. In: Proceedings of 14th International Conference on Software Engineering and Knowledge Engineering (SEKE), Ischia, Italy, pp. 351–358.

Fischer, G. (2001). “User Modeling in Human–Computer Interaction”. In: User modeling and user-adapted interaction, pp. 65–86.

García, F., Bertoa, M.F., Calero, C., Vallecillo, A., Ruíz, F., Piattini, M. and Genero, M. (2006). “Towards a consistent terminology for software measurement”. Information and Software Technology. pp. 631-644.

Girardi, R. and Faria, C. (2003). “A Generic Ontology for the Specification of Domain Models”. In: Proceedings of 1st International Workshop on Component Engineering Methodology (WCEM’03) at Second International Conference on Generative Programming and Component Engineering, Erfurt, Germany.

González-Pérez, C. and Henderson-Sellers, B. (2006). “An Ontology for Software Development Methodologies and Endeavours”. Ontologies for Software Engineering and Technology, Springer-Verlag, Berlin.

Hilera, J.R., Sánchez-Alonso, S., García, E. and Del Molino, C.J. (2005). “OntoGLOSE: A Light-weight Software Engineering Ontology”. In: 1st Workshop on Ontology, Conceptualizations and Epistemology for Software and Systems Engineering (ONTOSE), Alcalá de Henares, Spain.

Kitchenham, B.A., Travassos, G.H., Mayrhauser, A., Niessink, F., Schneidewind, N.F., Singer, J., Takada, S., Vehvilainen, R. and Yang, H. (1999). “Towards an Ontology of Software Maintenance”. Journal of Software Maintenance: Research and Practice, pp. 365–389.

Larburu, I.U., Pikatza, J.M., Sobrado, F.J., García, J.J. and López, D. (2003). “Hacia la implementación de una herramienta de soporte al proceso de desarrollo de software”. In: Workshop in Artifificial Intelligence Applications to Engineering (AIAI), San Sebastián, Spain.

Lin, S., Liu, F. and Loe, S. (2003). “Building A Knowledge Base of IEEE/EAI 12207 and CMMI with Ontology”. In: Sixth International Protégé Workshop, Manchester, England.

Mendes, O. and Abran, A. (2005). “Issues in the development of an ontology for an emerging engineering discipline”. In: First Workshop on Ontology, Conceptualizations and Epistemology for Software and Systems Engineering (ONTOSE), Alcalá de Henares, Spain.

Noy, N. F. and 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.

Ruiz, F., Vizcaíno, A., Piattini, M. and García, F. (2004). “An Ontology for the Management of Software Maintenance Projects”. International Journal of Software Engineering and Knowledge Engineering, pp. 323–349.

Ruiz, M., Diaz, M., Soler, F. and Perez, J. (2008). “Adaptation in current e-learning systems”. Computer Standards & Interfaces, pp. 62-70.

Sánchez, D.M., Cavero, J.M. and Marcos, E. (2005). “An ontology about ontologies and models: a conceptual discussion.” In: First Workshop on Ontology, Conceptualizations and Epistemology for Software and Systems Engineering (ONTOSE), Alcalá de Henares, Spain.

Sicilia, M., Cuadrado, J. J., Garcia, E., Rodriguez, D. and Hilera, J. R. (2005). “The evaluation of ontological representation of the SWEBOK as a revision tool”. In: 29th Annual International Computer Software and Application Conference (COMPSAC), Edinburgh, UK, pp. 26–28.

Stanford (2011). “The Protégé Ontology Editor and Knowledge Acquisition System”, http://protege.stanford.edu/index.html, April.

Tansalarak, N., Claypool and K.T. (2004). “XCM: A Component Ontology.” In: Workshop on Ontologies as Software Engineering Artifacts (OOPSLA), Vancouver, Canada.

Tautz, C. and Von Wangenheim, C.(1998). “REFSENO: A Representation Formalism for Software Engineering Ontologies”. Fraunhofer IESEReport No. 015.98/E, version 1.1, October 20.

Vizcaíno, A., Anquetil, N., Oliveira, K., Ruiz, F. and Piattini, M. (2005). “Merging Software Maintenance Ontologies: Our Experience”. In: First Workshop on Ontology, Conceptualizations and Epistemology for Software and Systems Engineering (ONTOSE), Alcala de Henares, Spain.

Yu, Z., Zhou, X. and Shu, L. (2010). “Towards a semantic infrastructure for context-aware e-learning”. Multimedia Tools and Applications, pp. 71–86.

Wongthongtham, P. (2006). “A methodology for multi-site distributed software development.” PhD Thesis, Curtin University of Technology.

Reasoning over visual knowledge*

Joel Luis Carbonera[1], Mara Abel[1], Claiton M. S. Scherer[2], Ariane K. Bernardes[2]

[1] Institute of Informatics – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[2]Institute of Geoscience – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

{jlcarbonera,marabel}@inf.ufrgs.br, {claiton.scherer,ariane.kravczyk}@ufrgs.br

Abstract. *In imagistic domains, such as Medicine, Meteorology and Geology, the tasks are accomplished through intensive use of visual knowledge, offering many challenges to the Computer Science. In this work we focus in an essential task accomplished in many imagistic domains: the visual interpretation task. We call visual interpretation the expert reasoning process that describes a cognitive path that starts with the visual perception of domain objects, involves the recognition of visual patterns in these objects and results in the understanding of the scene. We investigate the role played by foundational ontologies in problem solving methods involving visual information. We propose a cognitive model for visual interpretation that combines domain ontologies, ontologically well founded inferential knowledge structures based on the notion of perceptual chunks and PSM's. The proposed model was effectively applied through a Problem-solving method to solve the task of visual interpretation of depositional processes, within the Sedimentary Stratigraphy domain.*

1. Introduction

Imagistic domains are those in which the problem-solving process starts with a visual pattern-matching process, which captures the information that will further support the abstract inference process of interpretation. In this sense, imagistic domains make intensive use of *Visual Knowledge*, which is the set of mental models that support the process of reasoning over information that comes from the spatial arrangement and other visual aspects of domain entities [Lorenzatti et al. 2011]. Imagistic domains impose many challenges to Computer Science, in terms of acquisition, modeling, representation and reasoning, due to the tacit and unconscious nature [Polanyi 1966] of visual knowledge.

In the computational processing of visual data in imagistic domains, one aims to represent, extract and reason over the raw data, according to the meanings defined by the human mind. In this sense, we consider that the computational processing of visual data is a problem composed by several sub-problems. In general, the recent studies are focusing mainly in two of these sub-problems: the semantic representations of raw visual data [Lorenzatti et al. 2011] and the symbol grounding problem [Hudelot et al. 2005]. The former problem concerns to the development of computational representations that abstracts the raw visual data and captures the meaning of it, in a useful way for human beings. This is an important problem, since the meaning is established in human mind,

*This work is supported by the Brazilian Research Council (CNPq), CT-PETRO and ENDEEPER®

not in the visual data. The latter problem concerns to the issue of embodying the semantic interpretation of symbols into artificial systems, allowing it to establish the relation between the symbols and the raw visual data. We consider that there is a third problem that is addressed in the recent investigations as a less important one: the visual interpretation task, that is, the expert reasoning process that describes a cognitive path that starts with the visual perception of domain objects, involving the recognition of visual patterns in these objects and results in abstract conclusions which are meaningfully connected to these perceptions, that is, the understanding of the scene. According to [De Groot and Gobet 1996] "cognition is perception", in the sense that in the expertise development, the subjects develop dynamic abstractions of visual patterns of domain objects or visual features of domain objects, which guides the problem-solving process. We are interested in visual interpretation processes with these features.

The literature shows many approaches to deal with computational processing of visual data, such as low-level image processing, machine learning and knowledge-based approaches. Approaches that apply Image processing [Rangayyan et al. 2007] and machine learning [Akay 2009] techniques are based on detectable geometric features of the image (such as texture and shape) extracted from the raw data. These features cannot support the inferences that are developed in a more abstract level by the experts, as demonstrated in [Abel et al. 2005]. On the other hand, knowledge-based approaches, as semantic image interpretation, aims to model the abstract portion of knowledge that supports visual data understanding and to process this information symbolically, in order to reach conclusions in the domain. Recently, knowledge-based approaches make use of ontologies and Problem-Solving Methods (PSM) [Mastella et al. 2005]. Other works [Hudelot et al. 2005] propose the integration of image processing and knowledge-based approaches with reasoning capabilities, to interpret the raw visual data. However, most of the recent works have not been focused in the investigation of the human-like capabilities of reasoning over the visual knowledge abstracted from the visual data. Thus, the complete characterization of the inferential knowledge structures, needed to carry out visual interpretation tasks in a way that reproduces the human performance, is viewed as a secondary issue. In this work we address this issue, proposing an inferential knowledge structure for visual interpretation tasks.

We claim here that there are ontological meta-properties of domain concepts that provide the conditions which allow the visual perception of it instances, determining the domain concepts that can participate in the visual interpretation tasks. In this work we attempt to clarify the meta-properties of domain ontology primitives that allows visual interpretation tasks, exploring the role of foundational ontologies in the problem-solving methods used for this kind of task. This ontological clarification should allows the definition of inferential knowledge structures and PSM's that embodies ontological constraints of foundational ontology, increasing the potential reuse of them and allowing a more accurate mapping to the domain ontology. We propose here a knowledge-based computational approach for visual interpretation task that combines domain ontologies with an inferential knowledge structure, called *visual chunk*, which is based on the cognitive notion of perceptual chunk. The properties that the visually observable entities have in the point of view of the human cognition are reproduced in visual chunks through ontological constraints. In this sense, our approach relies on the meta-properties defined in the *Unified Foundational Ontology* [Guizzardi 2005] in order to establish the mapping between

the domain ontology and the inferential knowledge structures. Our approach offers some benefits: (a) approximates the class of possible inferential knowledge models to that of intended ones; (b) captures in a narrowest way the organization of the inferential knowledge used by the expert; (c) guides the process of acquisition of the inferential knowledge for visual interpretation tasks, and (d) helps to manage and to maintain the inferential knowledge in the systems.

The Section 2 presents the cognitive and technical foundations of our approach. Section 3 details our inferential knowledge structure for visual interpretation tasks. In this work we deal with a specific type of visual interpretation task, which concerns the visual interpretation of the events responsible by the generation of the visually observed object. Thus, we work with an instance of this task, that is, the visual interpretation of depositional processes responsible by the generation of sedimentary facies, in the domain of Sedimentary Stratigraphy. For these reasons, in section 4 we present an overview of the Sedimentary Stratigraphy domain and describe a PSM that applies the proposed approach in this domain. The section 5 presents the evaluation process of our approach and an analysis of the outcomes. Finally, section 6 presents our main conclusions.

2. Cognitive and technical foundations

We describe here the core theoretical framework of our work, including some studies of the human visual processing, the cognitive characterization of expertise, the unified foundational ontology (UFO) and PSM's.

2.1. Human visual processing

According to [Matthen 2005], the object perception depends on establishing a direct, causal and informational relation with a set of external physical objects, that corresponds to any unique material body that possesses hierarchically organized and cohesive parts, which exists independently of internal states of the perceiver and his/her perceptual systems. Moreover, in [Tversky 1989] it is pointed out that the notion of *parts* and *partonomies* play an important role in the perceptual processes. In this sense, the parts of a complex object play the role of perceptual saliencies, which provide important clues to individuate and recognize the object, through visual perception. The proper configuration of parts determines the shapes that objects can take. In addition, parts, and their perceptual saliencies, seem to be natural units of perception and natural units of function. In this sense, they provide important criteria in order to make more abstract judgments related to the perceived object, such as, functions and behaviors.

2.2. Expertise

The experts organize the knowledge in a qualitatively superior way, influencing the access to the knowledge and the interpretations of the perceptual stimuli coming of the environment. For the experts, the indexes of access to the knowledge are chunks of related perceptual stimuli that, when recognized together, allow the fast access to the knowledge meaningfully associated. These *perceptual chunks* are developed through the repeated recognition of the perceptual stimuli associated to specific situations or events, and play the role of cognitive triggers to the abstractions of those events and situations [Chase and Simon 1973]. Thus, perceptual chunks integrate sets of related perceptual

stimuli to more abstract conceptual components and can be seen as abstractions of a solution step in a problem-solving process [Cooke 1992]. In high degrees of expertise, the problem-solving process of visual interpretation tasks is driven by pattern-matching, where the visual stimuli that come from the domain are confronted with the visual patterns stored in perceptual chunks, triggering the abstract interpretations related to them. The gradual elaboration of perceptual chunks leads to the automation of the cognitive processes that integrate perception and the high-level cognition [Sternberg 1997]. These processes can explain the resistance that experts show in verbalizing the fine-grained knowledge that relates the visual aspects of the domain and their high-level interpretations. This intermediary knowledge, composed by domain explicit facts and rules of the domain theories in the early stage of expertise, is chunked and automated during many years of repeatedly application, integrating some specific knowledge developed during the practical activities, which transcends the explicit domain theories offered in the domain literature. Thus, perceptual chunks play the role of cognitive shortcuts, from the visual stimuli to the abstract interpretations related to them.

2.3. Unified Foundational Ontology

Foundational ontologies are meta-ontologies that have been developed based on the theories of a philosophical discipline called Formal Ontology. Foundational ontologies offer guides to make modeling decisions in the conceptual modeling process, clarifying and justifying the meaning of the models, improving the understandability and reusability. In this paper we use domain ontologies to represent the domain shared conceptualizations, and the Unified Foundational Ontology (UFO) to formalize both, the domain ontology and the inferential knowledge model. We will summarize the main UFO features that we will apply in this work. A full description of UFO can be found in [Guizzardi 2005]. UFO defines a set of meta-types and meta-properties that classify concepts in conceptual models. Initially, UFO makes a distinction between *Endurant Universal* and *Perdurant Universal* (or *Event Universal*). Instances of an *Endurant Universal* (such as Dog, Person, Country, etc) are individuals wholly present whenever they are present. On the other hand, instances of a *Perdurant Universal* (such as Game, War, etc), are individuals composed by temporal parts, that is, they happen in time, accumulating temporal parts. Within the *Endurant Universals*, UFO defines *Substantial Universals* whose instances are individuals that posses spatial-temporal properties, are founded on matter and are existentially independent from all other individuals. The relation between a *Substantial Universal* and an *Event Universal* is called *participation*, according to UFO. Some *Substantial Universals* are *Sortal Universals*, which provide principle of identity (PI) and principle of unity (PU). In this context, PI supports the judgment whether two instances of the universal are the same, when PU supports the counting of the instances of the universal. *Kind* is a *Sortal Universal* whose instances are functional complexes. On the other hand, *Moment Universals* are *Endurant Universals* whose instances are existentially dependent individuals that inheres in other individuals. Some *Moment Universals* are *Quality Universals*, which represents the properties in the conceptual models. A *Quality Universal* characterizes other Universals and it is related to *Quality Structures*, that is, a structure that represents the set of all values that a quality can assume. Thus, considering the property color as a *Quality Universal*, a given instance of Car could be characterized by an instance of quality Color, which is associated with a value of ColorStructure, which represents all the possible values that the property color can assume. Finally, UFO proposes four types of

parthood relations, clarifying their semantics: *componentOf*, *memberOf*, *subCollectionOf* and *subQuantityOf*. Each parthood relation only can be established between individuals of specific UFO meta-types, respecting some ontological constraints embodied in UFO.

2.4. Problem-solving methods

A PSM consists of an abstract specification that describes the reasoning process at the knowledge level, capturing the expert problem-solving behavior in a domain and implementation independent way, through the specification of the knowledge and control structures required [Perez and Benjamins 1999]. A reasoning pattern is modeled through a PSM, by three components: (i) a *competence specification* that describes what the PSM can do, (ii) an *operational specification* that describes how the process is developed and the knowledge required in each inference step of the process, and (iii) *requirements/assumptions* embodied in the method in terms of domain knowledge.

3. Inferential knowledge structures for visual interpretation tasks

Our approach adopt the notion of perceptual chunk in order to propose a structure of inferential knowledge representation that captures the direct relationship between the visual stimulus and the abstract interpretations meaningfully related to them, in a cognitively well founded way. Moreover, the inferential knowledge representation structure proposed here, called *Visual Chunk*, is organized as patterns of constrained arrangements of domain knowledge. This organization is the result of ontological constraints that allow the participation of only certain domain concepts and relations, arranged in specific way. In visual chunks, only instances of domain concepts classified as *Substantial Universal* according to UFO, can be visually perceived, since substantial universals have instances that satisfy the visual perception conditions: material bodies, which exists independently of internal states of the perceiver and his/her perceptual systems. Thus the core of a visual chunk is a Substantial Universal. The visual stimuli stored in the visual chunk, are values that belong to quality structures associated to quality universals, which characterizes the substantial universal, whose instances are visually inspected by the expert. Furthermore, our model preserves the importance of the parthood relations to the human perceptual and cognitive processes. We claim that an effective modeling of inferential knowledge structures and inference processes in imagistic domains should be focused in revealing and representing the perceptual chunks applied by experts, avoiding the problems related to elicit the tacit fine-grained knowledge that relates perceptual stimuli and their abstract interpretations.

3.1. Characterization

Let O be a domain ontology, V is the vocabulary that represents this ontology. The vocabulary of interest to the realization of the task of visual interpretation of events is denoted by V_{target}, and corresponds to a subset of V. The V_{target} contains two pairwise disjoint subsets: V_{vk} e V_{int}. The V_{vk} represents the domain primitives (concepts, relations and properties) used by the expert to describe visually the objects of interest in the domain. While V_{int} corresponds to the vocabulary that represents the domain primitives that describe events that can be interpretable through visual inspection of the domain objects described by V_{vk}. Thus:

$$V_{target} \subseteq V$$

$$V_{target} = V_{vk} \cup V_{int}$$
$$V_{vk} \cap V_{int} = \emptyset$$

In a very abstract level, a *Visual Chunk* has the general form of a logical implication, such as

$$antecedent \Longrightarrow consequent,$$

where the $antecedent$ is a logical formula, constituted by atoms a_{vk}, where $a_{vk} \in V_{vk}$, and the $consequent$ is a logical formula, constituted by atoms a_{int}, where $a_{int} \in V_{int}$. Our aim is to restrict the vocabularies V_{vk} and V_{int}, considering UFO ontological constraints to reflect the cognitive constraints previously discussed. In this sense, these vocabularies can represent only certain meta-concepts and relations offered by UFO.

The vocabulary V_{vk} must contain only and exclusively the following constructs:

ObservableEntity: Represents domain primitives whose instances can be directly visually perceived. We consider that only instances of domain concepts classified as *Substantial Universal*, according to UFO, can be direct visually perceived.

VisualQuality: Represents the abstraction of a possible visual quality of a domain entity visually observable. In this sense, according to UFO, they are *Quality Universals* defined in the domain ontology, which maintains a *characterization* relation with an *ObservableEntity*.

VisualQualia: Represents a constrained set of possible values of a *VisualQuality*. Is a subset of values that belong to the *Quality Structure* associated to a *VisualQuality*.

VisualQuale : Represents a value that belongs to the *Quality Structure* associated to a *VisualQuality*.

PartOfRelation: Represents a parthood relation between two *ObservableEntity* in the domain. This relation is one of that allowed by UFO. The specific type of parthood relation depends on the specific ontological nature of the two *ObservableEntity* related, following the ontological restrictions imposed by UFO.

The vocabulary V_{int} must contain only and exclusively the following constructs:

InterpretableEvent: Represents domain concepts that abstract the events responsible by the generation of the *ObservableEntity*. These concepts are classified as *Event* in the UFO. As an additional requisite, these concepts must be organized in a subsumption hierarchy, since the interpretation task aims to find the more specific subtype of *InterpretableEvent* responsible by the generation of the *ObservableEntity* individual under visual inspection.

ParticipationRelation: Represents a domain *participation* relation between the *ObservableEntity* whose instance is being interpreted, and the *InterpretableEvent* responsible by it generation.

The *Visual Chunk* is structured according to some internal structures, which represent recurrent patterns of relationship among the constructs previously presented. This structure can be described as following, in a semi-formal way. Firstly, a *VisualChunk* is the structure that relates *VisualFeatures* and an *Interpretation*.

$$VisualChunk =_{def} (VisualFeatures, Interpretation)$$

VisualFeatures can be simple (*SimpleVisualFeatures*) or complex (*ComplexVisualFeatures*).

$$VisualFeatures =_{def} SimpleVisualFeatures \vee ComplexVisualFeatures$$

SimpleVisualFeatures is a structure that relates an *ObservableEntity* and a set of *PossibleVisualFeatures*.

$$SimpleVisualFeatures =_{def} (ObservableEntity, \{PossibleVisualFeatures_1, ..., PossibleVisualFeatures_n\})$$

where

$$VisualFeatures \Longrightarrow Interpretation$$

PossibleVisualFeatures is a structure that relates a *VisualQualia* and a *VisualQuality*, which maintains a *characterization* relation with the *ObservableEntity*. Here, *VisualQualia* corresponds to a constrained sub-set of values of the *Quality Structure* associated to the *VisualQuality* in the domain ontology. This sub-set of values represents the values that the *VisualQuality* can assume to support the *Interpretation* according to the expert.

$$PossibleVisualFeatures =_{def} (VisualQuality, VisualQualia)$$

ComplexVisualFeatures, on the other hand, is a structure that relates a *SimpleVisualFeatures* to a set of *VisualPart*.

$$ComplexVisualFeatures =_{def} (SimpleVisualFeatures, \{VisualPart_1, ..., VisualPart_1\})$$

VisualPart is a structure that relates a *PartOfRelation* to a set of *VisualFeatures* derived from ($ObservableEntity_{part}$), which are parts of the *ObservableEntity* (representing the whole visually observed). The *PartOfRelation* relates the *ObservableEntity* that are wholes to the $ObservableEntity_{part}$, which are their parts.

$$VisualPart =_{def} (PartOfRelation, \{VisualFeatures_1, ..., VisualFeatures_n\})$$

Interpretation is a structure that relates a *ParticipationRelation* to an *InterpretableEvent*.

$$Interpretation =_{def} (ParticipationRelation, InterpretableEvent)$$

In this sense, considering a specific *VisualChunk*, when *VisualFeatures* is *found*, then *Interpretation* is also *found*. In the case of *VisualFeatures* to be a *SimpleVisualFeatures*, we say that it is *found* when all the *PossibleVisualFeatures* related to the *ObservableEntity* are *found*. A *PossibleVisualFeatures* is *found* when there is a *Quality individual* that is instance of the *VisualQuality*, which *inheres in* the particular *ObservableEntity* under visual inspection and that assumes a value which is a *VisualQuale* that belongs to the correspondig *VisualQualia*. On the other hand, a *ComplexVisualFeatures* is said *found* when the *SimpleVisualFeatures* is *found* and all the *VisualPart* are *found*. A *VisualPart* is *found* when there is a *PartOfRelation* between the *ObservableEntity* and an $ObservableEntity_{part}$ that is its part, and when at least one of the *SimpleVisualFeatures* (derived from the $ObservableEntity_{part}$) is *found*. When the *Interpretation* is *found*, the *ParticipationRelation* that relates the *ObservableEntity* and the *InterpretableEvent* is instantiated.

4. Case study: Sedimentary Stratigraphy

Sedimentary Stratigraphy is the study of sedimentary terrains in surface or subsurface of the Earth, in order to define the geological history of their formation based on the visual

description of well cores and outcrops. The main objects of study and description is: Sedimentary Facies (SF), Sedimentary Structures (SS) and Depositional Processes (DP). A SF is a region in a well core or outcrop, visually distinguishable of adjacent regions. Each SF is assumed as a direct result of the occurrence of a DP. A SS is the external visual aspect of some internal spatial arrangement of the rock grains. Finally, DP are events that involve the complex interaction of natural forces and sediments. DP are responsible for the formation of sedimentary rocks, through transport and deposition of sediments in a sedimentation place. Our domain ontology of Sedimentary Stratigraphy is ontologically well founded, using UFO. In this work, we present the ontological characterization of these three main concepts, but the properties will not be fully detailed.

Sedimentary Facies (SF): Instances of SF can be visually recognized, individuated and counted. SF offers a principle of identity and it instances cannot cease to be SF without ceasing to exist. According to UFO it is a *Kind*. The set of *Quality Universals* that characterizes SF include: lithology, sorting, roundness and others. There is a relation between SF and SS, called *hasSedimentaryStructure*, which is a *componentOf* relation, according to UFO.

Sedimentary Structures (SS): It is an analogous case to SF. Therefore it is also a *Kind*. The SS concept has many subkinds organized in a taxonomy. The set of *Quality Universals* that characterizes SS include: laminae shape, angularity, thickness, laminae shape and so on.

Depositional Process (DP): Entities of DP happen in time. We consider DP as an *Event*. Since the SF's are the final results of a DP occurrence, they are participants of DP, that is, there is a *participation* relation, called *generatedBy*, between SF and DP. There are many specific types of DP, which are organized in a taxonomic structure.

During the inspection of a well core or an outcrop, the expert visually segments the body of rock in many distinct SF, observing several discontinuities of the visual properties. After this segmentation process, each SF is visually examined to interpret a specific type of DP, since that each SF was generated by a DP occurred in a remote past. The expert points out the DP by visually observing an aggregation of visual stimuli of the rock, which preserves many visual features which record the action of plastic forces of the DP occurrence. This interpretation process is based on the expert extensive previous knowledge, indexed by perceptual chunks. Thus, the elicitation of these perceptual chunks was a core question of the interaction with the domain expert, during the knowledge acquisition process. The Figure 1 presents an instance of *Visual Chunk* built on the domain ontology of Sedimentary Stratigraphy.

The reasoning pattern that the expert uses to interpret visually Depositional Processes was abstractly captured in a PSM (represented in Figure 2). The PSM uses the *Visual Chunks* presented in the section 3, as inferential knowledge structures. The *competence* of our PSM takes a visual description of an *Observable Entity* and a taxonomy of *Interpretable Events* and infers the specific *Interpretable Event* indicated by the *Observable Entity*. The *assumption* of our PSM is that the visual features imprinted in the *Observable Entities* of the domain indicate an *Interpretable Entity*. The *requirements* are the visual chunks that the expert applies to relate the visual stimuli of the *Observable Entities* to *Interpretable Event*. The *operational specification* describes the inferences in the PSM, which can be detailed as follow:

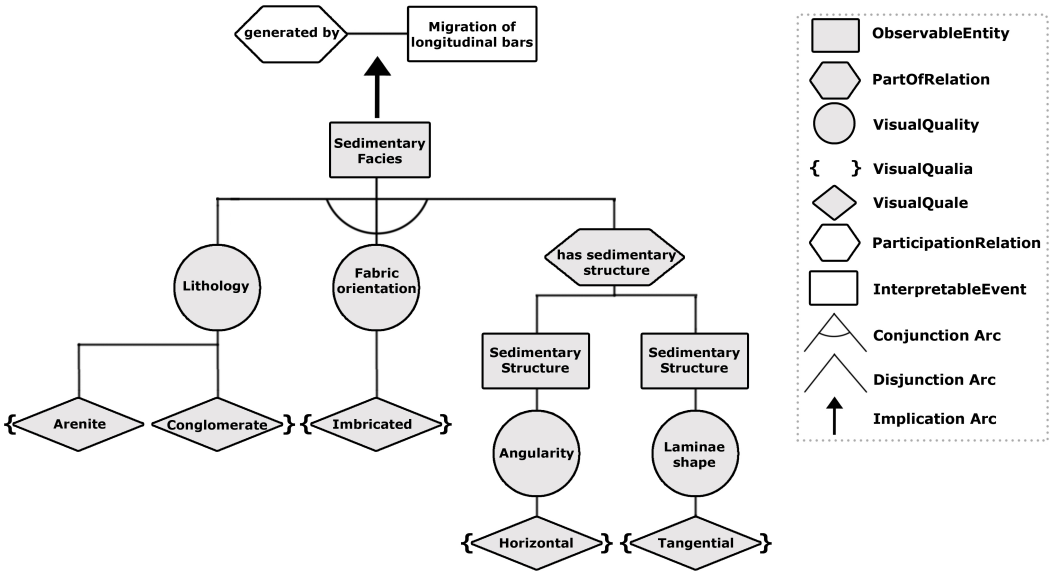


Figure 1. Representation of a Visual Chunk for interpretation of Depositional Processes

Generate: Generate *candidate* interpretations according to the constraints of the taxonomy of *Interpretable Entities* in domain ontology.

Retrieve: Retrieves a set of *Visual Chunks*, whose *Interpretable Event* corresponds to the current *Candidate* interpretation.

Select: Selects a *Visual Chunk* of the previously retrieved set of *Visual Chunks*.

Decompose: Decomposes the *Observable Entity* in other *Observable Entity* that compose it.

Specify: Specifies relevant *Visual Attributes* of the *Observable Entity* and its components.

Obtain: Obtains relevant *Visual Features* (visual attributes and values assigned to then).

Match: Tries to match a specific *Perceptual Chunk* to the relevant features of the *Observable Entity*.

Assign: Assigns the *Candidate* as the current interpretation, in case of positive match of *Visual Features* and *Perceptual Chunk*.

The PSM receives as input an *Observable Entity* (5) and *Interpretable Entities* (1) organized in a taxonomy. The *Observable Entity* is decomposed in other *Observable Entities* that compose it (8). Relevant visual attributes (6 and 9) of the *Observable Entities* are specified, and visual features are obtained (7 and 10) from these attributes. Candidate interpretations (2) are generated from the taxonomy of *Interpretable Entities* (according to the subsumption hierarchy in the domain ontology). A set of *Visual Chunks* (3) whose *Interpretable Event* corresponds to the current candidate interpretation is retrieved. From this set, it is selected a *Perceptual Chunk* (4). Finally, the PSM tries to match the *Perceptual Chunk* with the *Visual Features* of the *Observable Entity*. In the case of positive match (11), the candidate is assigned as the current interpretation (12). This process traverses the taxonomy of *Interpretable Entities* in a top-down way, trying to reach a more specific interpretation in each step. The final interpretation is the last *Interpretable Entity* with at least one *Perceptual Chunk* matched. The process occurs until a leaf of the

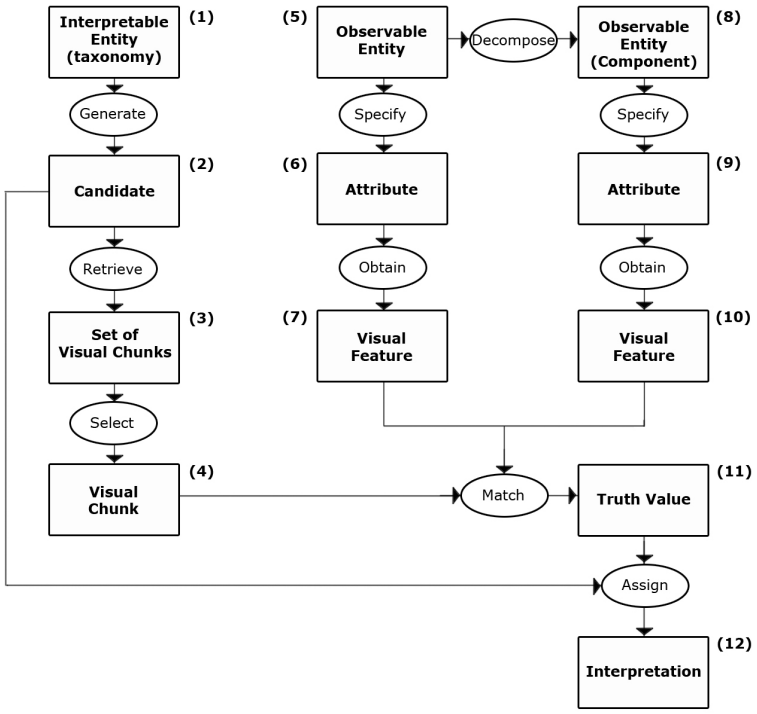


Figure 2. Representation of the PSM to visual interpretation of Depositional Processes

taxonomy is reached or the matching process return false to all the *Perceptual Chunks* associated to the candidate interpretation (meaning that only an interpretation of intermediary level of specificity can be reached from the visual description at hand). Thus, this reasoning model can be viewed as a process of hypothesis generation, retrieval of its associated *Visual Chunks* and symbolic pattern matching between the *Visual Chunks* and the symbolic visual description of the domain objects. The hypotheses are generated according to the constraints embodied in the taxonomy of *Interpretable Entities*. In this sense, when a hypothesis is proved, the next step assumes the concepts of the next level of the taxonomy as the new set of hypothesis to be tested.

5. Evaluation of the approach

Our approach was applied to interpret a set of three real stratigraphic descriptions available in the literature, and interpretations carried out by our approach were compared with the interpretations offered by the literature. Since that this work was focused only in a sub-type of depositional processes called *depositional processes of tractive currents*, it is expected that the PSM interprets only sedimentary facies that had been generated by processes of this sub-type. In other cases, it is expected that the PSM interprets the process as a *Depositional Process*, the more general process in the taxonomy. We consider that an outcome of this type is an *inconclusive interpretation* (a non-answer). Since the PSM can reach interpretations in several levels of generality/specificity, is also expected that for some cases, the conclusion generated by the PSM will be more general than the interpretation of the literature. Thus, to evaluate our approach in detail, we defined some distinct categories of outcomes. Firstly, the outcomes can be satisfactory, when the outcome of the PSM is compatible with the expected interpretation; or unsatisfactory, when the outcome

is incompatible with the expected interpretation. Among the *unsatisfactory outcomes*, we distinguish the *false negatives*, when the outcome is an *inconclusive interpretation* and an interpretation was expected; and *false positives*, when the PSM had offered an interpretation and was expected an *inconclusive interpretation*, or when the outcome is a specific interpretation that do not corresponds to the expected interpretation. Within the *satisfactory outcomes*, we distinguish the *true negatives*, when the outcome is an *inconclusive interpretation* for the cases in that the depositional process is not a *depositional process of tractive currents*; and the *true positives*, when the outcome is compatible with the expected interpretation. Finally, within the *true positives* we distinguish the *Specific* correspondences, when the approach provides the more specialized interpretation according to the input; and *General* correspondences, when the outcome is a generalization of the expected interpretation. The Table 1 shows an analysis of the evaluation process.

Table 1. Analysis of the outcomes of the evaluation process

Evaluated cases	Number of facies	Unsatisfactory Outcomes		Satisfactory Outcomes		
		False positives	False negatives	True negatives	True positives	
					Specific	General
Case 1	14	0%	0%	50%	36%	14%
Case 2	8	0%	0%	50%	38%	12%
Case 3	7	0%	0%	29%	57%	14%

The evaluation analysis showed that, for the considered datasets, all the results accomplished had been satisfactory. However the analysis also revealed that, for a significant percentage of facies descriptions, our approach offered interpretations more general than those offered by the literature. One hypothesis that explain this observation is the possibility of the visual descriptions of datasets to be excessively general to support the specificity of the interpretations offered in the literature. This hypothesis will be investigated in future works.

6. Conclusion

We described a modeling approach to explicitly deal with the semantic embedded in visual objects that are used by experts to support problem solving. We built our approach based on the comprehension about how people individuate significant objects when scanning them through the visual system. We recognized that the notion of perceptual chunk, previously identified in several studies, plays a fundamental role in the connection of perceptual capture and further interpretation inference over the domain knowledge. Therefore, we showed that the inherent properties of visually recognized objects can be identified and expressed using constructs that are ontologically founded. The ontological constructs provide the necessary independence between the application and the model that allows reusing both, the reasoning algorithms and the domain ontology. Thus, this work shows the role played by foundational ontologies in problem solving methods involving visual information. We have applied the proposed model to build a robust representation of visual knowledge in a complex real application in Petroleum Geology, and explore it to extract useful stratigraphic interpretations of events and their register in the Earth.

References

Abel, M., Silva, L. A., Campbell, J. A., and De Ros, L. F. (2005). Knowledge acquisition and interpretation problem-solving methods for visual expertise: study of petroleum-reservoir evaluation. *Journal of Petroleum Science and Engineering*, 47:51–69.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2):3240–3247.

Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1):55–81.

Cooke, N. J. (1992). *The Psychology of Expertise: Cognitive Research and Empirical AI*, chapter Modeling human expertise in expert systems, pages 29–60. Springer Verlag, New York.

De Groot, A. D. and Gobet, F. (1996). *Perception and memory in chess: Heuristics of the professional eye*. Van Gorcum, Assen.

Guizzardi, G. (2005). *Ontological Foundations for Structural Conceptual Models*, volume 05-74 of *CTIT PhD Thesis Series*. Universal Press, Enschede, The Netherlands.

Hudelot, C., Maillot, N., and Thonnat, M. (2005). Symbol grounding for semantic image interpretation : from image data to semantics. In *Proceedings of the Workshop on Semantic Knowledge in Computer Vision, ICCV*.

Lorenzatti, A., Abel, M., Fiorini, S. R., Bernardes, A. K., and dos Santos Scherer, C. M. (2011). Ontological primitives for visual knowledge. In *Proceedings of the 20th Brazilian conference on Advances in artificial intelligence (2010)*, volume 6404 of *Lectures Notes in Artificial Intelligence*, pages 1–10, São Bernardo do Campo. Springer Berlin / Heidelberg.

Mastella, L. S., Abel, M., Lamb, L. C., and De Ros, L. F. (2005). Cognitive modelling of event ordering reasoning in imagistic domains. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 528–533, Edinburgh, UK. Morgan Kaufmann Publishers Inc.

Matthen, M. (2005). *Seeing, Doing, and Knowing: A Philosophical Theory of Sense Perception*. Oxford University Press.

Perez, A. G. and Benjamins, V. R. (1999). Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods. In *Proceedings of IJCAI-99 Workshop on Ontologies and Problem Solving Methods (KRR5)*, Stockholm, Sweden.

Polanyi, M. (1966). *The tacit dimension*. Anchor Day Books, New York.

Rangayyan, R. M., Ayres, F. J., and Desautels, J. L. (2007). A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344:312–348.

Sternberg, R. J. (1997). Cognitive conceptions of expertise. In Feltovich, P. J., Ford, K. M., and Hoffman, R. R., editors, *Expertise in context*, chapter Cognitive conceptions of expertise, pages 149–162. AAAI/MIT Press, Menlo Park, California.

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, 25:983–995.

Ranganathans Canons applied to ontology engineering: a sample application scenario in biomedical ontologies

Linair Maria Campos[1], Maria Luiza de Almeida Campos[2], Maria Luiza Machado Campos[3]

[1] Universidade Federal do Rio de Janeiro – CISI/COPPE
Av. Horacio Macedo S/N - Ilha do Fundão – Bloco H-201- Rio de Janeiro - RJ

[2] Universidade Federal Fluminense - PPGCI/UFF
Rua Tiradentes 148 - Ingá - Niterói - Rio de Janeiro - RJ

[3] Universidade Federal do Rio de Janeiro - PPGI- IM/NCE
Av. Athos da Silveira Ramos, 274 - Ilha do Fundão - Rio de Janeiro - RJ

linair@cisi.coppe.ufrj.br, maria.almeida@pq.cnpq.br, mluiza@ufrj.br

Abstract. *Ranganathan, an Indian mathematician and librarian, has proposed a set of comprehensible canons to provide guidance to the process of building concept hierarchies. It is our proposition that Ranganathans canons can contribute to fulfill the gap between the high-level domain conceptualization guided by top level ontologies and the classification of such concepts within facets, needed when building ontologies taxonomical structures. In order to show the utility of Ranganathans canons applied to ontology structuring, we have analyzed the structure of a biomedical ontology: Gene Ontology (GO). As result, we have found that many of the existing inconsistencies on GO hierarchies could be avoided if Ranganathans canons were adopted.*

1. Introduction

Ontologies have been increasingly used since the early 90s, especially in complex domains such as Biomedicine, where the multitude of concepts and the need to deal computationally with resources described by them, has urged the adoption of standard vocabularies. However, the fast growing nature of the body of knowledge being described and the necessity of fast solution to the issue of concepts standardization has given rise to vocabularies such as the Gene Ontology [Gene Ontology Consortium, 2001], which has been created without a sound methodology and has been largely adopted as a *de facto* standard, despite its many structural problems, as mentioned in literature [Smith e Kumar, 2004][Smith, Williams, e Schulze-Kremer, 2003], which affects its efficient utilization.

Notions underlying concepts nature, materialized in classes of top level ontologies, have given significant contribution to ontology structuring, as it allows domain concepts to be identified and grouped together in basic categories according to pre-defined basic features. These top level classes are usually chosen according to principles discussed in areas such as Philosophy, Cognitive Sciences and Psychology, providing a sound basis for identifying the nature of concepts in a less ambiguous way. However, once groups of concepts with the same nature are identified, there is still the

need to organize them in arrays (horizontal series of sibling concepts) and chains (vertical series of concepts, organized hierarchically): this is one of the challenges of ontology structuring.

Underpinned by more than fifty years of hands on experience in information classification and structuring of big vocabularies, Shialy Rammarita Ranganathan, an Indian mathematician and librarian, has proposed a set of principles, or *canons,* [Ranganathan,1951;1963;1967a, 1967b], tailored to provide guidance to the process of working on the *level of ideas*. This level is the space where the concepts of a given domain are organized, building a system of concepts [Campos e Gomes, 2008]. These canons were meant to be used for bibliographic classification, in the context of the development of documentary languages with taxonomic structures such as thesauri and controlled vocabularies.

It is our proposition that Ranganathans' canons can contribute to fulfill the gap between the high-level domain conceptualization guided by top level ontologies and the systematic classification of such concepts within facets, needed when building ontologies taxonomical structures. In order to show the utility of these canons applied to ontology engineering, we have applied a set of those canons to the widely adopted, *de facto* standard, Gene Ontology (GO) [Gene Ontology Consortium, 2001].

As a result, we have observed the timeliness and relevance of his work, as a series of existing inconsistencies on GO hierarchies, could be avoided if Ranganathans canons were adopted.

The remainder of this article is structured as follows: in section 2 we present related work. In section 3 we discuss Ranganathans' canons. In section 4 we analyze the Molecular Function branch of GO in accordance with Ranaganathans canons. Finally, in section 5 we present our conclusions.

2. Related work

Ontology structuring has impact on knowledge reasoning, especially when based on subsumption relations. Reasoners expect ontologies to comply with certain rules of classification, and ill-formed hierarchies can lead to false results. On the other hand, implicit structuring strategies used to form subhierachies can hinder human comprehension of ontology classification rationale leading to ambiguity when using and extending ontologies.

In order to tackle those issues, researchers such as Guarino and Welty [2004] have proposed the use of philosophical notions, such as rigidity, identity, and unity, materialized on top ontologies, to guide the identification of concepts nature and, thus, to provide foundational principles to evaluate the conceptual correctness of specialization relationships [Guizzardi, 2005]. In this sense, Guarino and Welty have used those notions to underpin the *OntoClean methodology* [Guarino & Welty, 2004, 2002a, 2002b] aiming at building "clean" taxonomical structures on ontologies. Inspired by the work of Guarino and Welty, Guizzardi also has proposed a theory presenting a set of postulates aimed to aid the construction of well grounded conceptual models [Guizzardi, 2005; Guizzardi, Wagener and Sinderen, 2004]. The idea behind these approaches is to axiomatize a set of rules that can be applied systematically to taxonomies and on doing so, prevent structural errors. For example, based on the axiom

that rigid concepts cannot be subsumed by anti-rigid concepts, the concept `human` (rigid) cannot be subsumed by the concept `student` (anti-rigid).

Smith (2005) observes, complementary, the role of definitions to identify attributes "in a consistent manner, thus assuring their transitive inheritance through a type hierarchy", and points that the definition of a concept within an ontology should encompass the definition of all its parents. Besides, all intermediate classes in the hierarchy where the concept is situated should also be defined, in order to ensure transitive inheritance of essential characteristics. With the intention of establishing guidance for building well formed hierarchies, built according to a sound classification systematic, Smith proposes a set of axioms for a "Theory of Biological Classification". According to Smith (2005), those axioms were motivated by the theory of classes found in Aristotle's writings. As an example, we can mention the axiom that addresses the issue of polihierarchy and states that a species should never have two parents:

$$\text{lowestspecies}(A) \wedge \text{lowestspecies}(B) \wedge A \neq B \rightarrow \neg\exists x(\boldsymbol{inst}(x, A) \wedge \boldsymbol{inst}(x, B))$$

In Information Science, Dahlberg (1978a, 1978b) also stresses the importance of definitions as they make explicit the contents of concepts and provide the elements that forge the relationships between them. Dahlberg, through her Concepts Theory, proposes also that definitions reveal a set of common characteristics which are useful to build any system of classification or thesaurus [Dahlberg, 1983]. Dahlberg, however, focuses on proposing principles to organize concepts in broad categories, and, although highlighting with examples the importance of definitions when structuring hierarchies, she does not provide detailed guidance on how to organize them systematically in subclasses.

To exemplify the problems caused by the lack of a systematic approach on structuring an ontology taxonomy, Smith points out several issues in the Gene Ontology (GO) hierarchies. Although his axioms can help to identify solutions to those problems, members of the GO community were not very receptive to the proposal, perhaps due to the complexity that comes with it: "(…) When challenged with such problems, the members of the GO and associated communities standardly insist that their concerns are those of practicing biologists, and that they are thus not concerned with the sorts of scrupulousness that are important in logic" [Smith, 2005].

Guarino and Weltys' (2004) as well as Smiths' (2005) proposals have the focus on identifying concepts nature and on providing rules and axioms to help the identification and grouping of concepts with same nature in a consistent way. The inspiration behind the idea of identifying concepts nature, which has its roots in Philosophy, has been used since the 60s in the context of library classification by Ranganathan, who also adopted fundamental categories to help to identify and group vocabulary concepts according to their high level nature. In this context, Ranganathans' categories provide a more intuitive, transparent, although less formal (and consequently more ambiguous) way to approach categorization.

It is worth noting that if, for one hand, the use of formal axioms can improve ontology structuring, when applied by ontologists with some expertise in logics and with some background in Philosophy, on the other hand it can represent a challenge for

domain experts to deal with the inherent complexity of such philosophical notions and the formalisms used to express them [Yu, 2006].

However, although the identification of concepts nature has a major role in structuring well formed and consistent hierarchies, there is more to it than that. There is still the need of detailed classification principles to help organize concepts in subclasses, and, besides, if possible, that those principles could be more easily assimilated by the community responsible for creating and maintaining the ontologies. In this sense, to the best of our knowledge, few proposals provide a systematic set of principles to address the problem. Even so, some are directed specifically to a given subject, such as *folk biological classification of organisms* [Berlin, Breedlove, Raven, 1973] or *construction works* [ISO DIS 12006-2, 1999], while others [Ekholm, 2002] are focused on identifying objects properties, which even though helps on identifying facets of interest, does not present a solution to the issue of organizing them in a more thorough way.

Bodenreider and others (2004) point that there are some principles of good classification that (biomedical) ontologies are expected to be compliant and that, as they believe, "rest on a wide consensus among those working on biomedical terminologies". Such principles can be summarized as: (i) each hierarchy must have a single root; (ii) children should have exactly one parent; (iii) non-leaf classes must have at least two children; (iv) each class must differ from another class in its definition. In particular, each child must differ from its parent and siblings must differ from one another. The authors, however, do not present evidences on how long their proposal has been adopted and how exactly the consensus was reached. Besides, proposing that a hierarchy must have a single root seems to limit the possibility to express different aspects of a domain, which, in a different perspective, could be easily presented as facets. Also, their proposal does not provide guidance to other important aspects of classification, such as the need to define homogeneous hierarchies, as pointed out by Smith (2005).

As observed, although there has been some concern with the adoption of systematic classification practices, preventing *ad hoc* built taxonomies, many of those practices seem to be recent and still need to mature. Also, some of them lack a more intuitive and transparent explanation, in order to allow a better understanding by end users and so, to avoid their rejection.

It is our proposal that Ranganathans' canons of classification, in use for more than fifty years, provide a methodological path that can join the convenience of a comprehensive explanation and a more complete and mature set of guidelines, which can be easily adopted to help building more consistent classificatory structures. The usefulness of these canons can be seen in a sample scenario for analyzing Gene Ontology main classificatory structure, as presented bellow.

3. Ranganathans' canons

In the present paper, we highlight two sets of Ranganathans' canons, which provide guidelines to the organization of classes of concepts: canons for the creation of arrays and canons for the creation of chains. Chains are vertical series of concepts, which can be organized hierarchically according to generic-specific relations, or according to part-

of relations. Arrays are horizontal series of concepts, organized as siblings in relation to a parent concept.

In the specific case of the organization of ontologies taxonomical structures, we have selected a subset of Ranganathans' canons, of particular relevance to our purposes, and which we shortly present in the following sections, based on Campos e Gomes (2008) and also Gomes Motta e Campos (2006).

Some of the canons aim at organizing arrays, as for instance, the canons of Differentiation, Concomitance and Exclusivity, while others aim at organizing chains, as, for instance, the canons of Modulation and Subordinate Classes (Ranganathan, 1967a). The canons provide principles that facilitate the creation of classes in a more consistent way, and, according to Ranganathan (1967a), their violation may result in ill-formed classificatory structures. The selected canons are explained next.

3.1. Canons for organizing arrays

Ranganathans' **Differentiation canon** states that a principle of division used as a classificatory basis should originate at least two classes. For example, let us consider the array used to classify catalectic activities of enzymes. That array can have a principle of division according to the kind of enzymes (hydrolise, isomerase, among others) and another principle of division according to the kind of reaction catalyzed by the enzyme (free radical formation, first spliceosomal transesterification, among others).

If the principles of division used to organize the arrays are explicit, it makes the classification of new concepts easier, as the comprehension of the rationale used to form the hierarchy helps to figure it out where is the right place for the concept within the ontology structure:

> (...) in a classificatory scheme, concepts that are subordinated to a more general concept can be grouped more accurately according to the principle of division that guided this grouping. Principles of division bring transparency to the vocabulary and so improve searches, locating and relating the concept according to its inner characteristics. [Novellino, 1996, p.1].

Ranganathans' **Concomitance canon** states that two different principles of division should not result in the same array. For example, if we adopt the criteria of year of birth and age to classify a set of individuals, we will have as a result arrays constituted by the same elements.

Ranganathans' **Exclusivity canon** states that elements belonging to an array should be mutually exclusive, i.e., disjoint in relation to elements belonging to another array. For example, the term `multidrug transporter activity` should not be subordinate to both arrays `transmembrane transporter activity` and `drug transporter activity`. Even if those arrays are organized according to different principles of division (in the above example, according respectively to the principle of the local – `transmembrane` – where the transport occur and according to the kind of element – `drug` – which is transported).

3.2. Canons for organizing chains

As classificatory principles for chains, we highlight the following canons of Ranganathan: Modulation and Subordinate Classes [Ranganathan, 1967a] as explained next.

The **Canon of Modulation** states that within a hierarchical classificatory structure of concepts there should be a gradual specificity when organizing concepts in chains, allowing thus a "conceptual consistence between the classes of concepts" [Gomes, Motta e Campos, 2006]. For example, let us consider the terms `helicase`, `ATP-dependent RNA helicase` and `ATP-dependent DNA helicase`. According to the Canon of Modulation, the last two terms should not be directly subordinated to the term `helicase`. It should exist, between the first and the last two terms, a term like `ATP-dependent helicase`.

The **Canon for Subordinate Classes** states that in a hierarchy of classes, the classes nature should be the same, i.e., they should conform to the perspective adopted as principle of division that guides the organization of the array. For example, in GO, the definition of the term `binding` indicates that the principle of division of its array has to do with interaction between molecules. This makes us believe that terms like `bacterial binding` (Interacting selectively and non-covalently with any part of a bacterial cell) and "`extracellular matrix binding`" (Interacting selectively and non-covalently with a component of the extracellular matrix) fall in conflict with such principle, for a *bacterial*, which is an organism, and an *extracellular matrix*, which is a cellular component, both have different natures from "interaction between molecules", which is a process. According with Gomes, Motta e Campos (2006), this canon complements the Canon of Modulation, and, if it is not violated, the affiliation sequence of the chain to the array is correctly assured.

4. Analyzing GO (Molecular Function) and observing Ranganathans' canons

Our considerations about the utility of Ranganathans' canons are made in the context of the analysis of the *ontological commitment* of the Molecular Function branch of GO.

Ontological commitment can be defined briefly as an agreement shared by a community about the consensual meaning intended for the ontology, not only considering its comprehension by humans, but also considering its computational processing by software agents. We assume that this commitment is not always precisely explicit, however it can be identified, although partially, by means of the existing ontology documentation, the analysis of concepts definition and the metadata associated to ontologies terms. On retrieving the ontological commitment, we expect to have as results the criteria observed as classificatory principle for the organization of first level hierarchies of GO's Molecular Function branch, together with the problems observed as consequence of the adopted classificatory approach.

The choice of GO´s Molecular Function branch was due to the fact that this branch has intermediary complexity, if compared to GO´s Cellular Component (less complex) and Biological Process (more complex), as we could notice when analyzing the terms definition and also the composition of the first level hierarchies (considering

the hierarchies' depth, number of terms and existing relationships). In this sense, this choice is adequate to our purposes, for if, on one hand, it provides richness of issues to explore, on the other hand it minimizes the complexity, already big, of the analysis of the ontological commitment of GO.

The analysis of the ontological commitment is made upon the analysis of the ontology's hierarchical structure, the terms nomenclature, and especially of the terms definition, which is of main relevance to the formation and comprehension of domains classificatory structures, once they provide rich semantics about intended meaning of concepts.

For the sake of space, however, we are not going to reproduce the definition of subordinate terms, but only data about the first level term being analyzed, and the final result obtained, i.e., the criteria observed as classificatory principle, along with the summary of problems found, considering the array being analyzed. In order to help understanding the analysis of the `binding` array, we present on Figure 1 its immediate subordinate classes. The `binding` array contains more than 1000 subordinate classes.

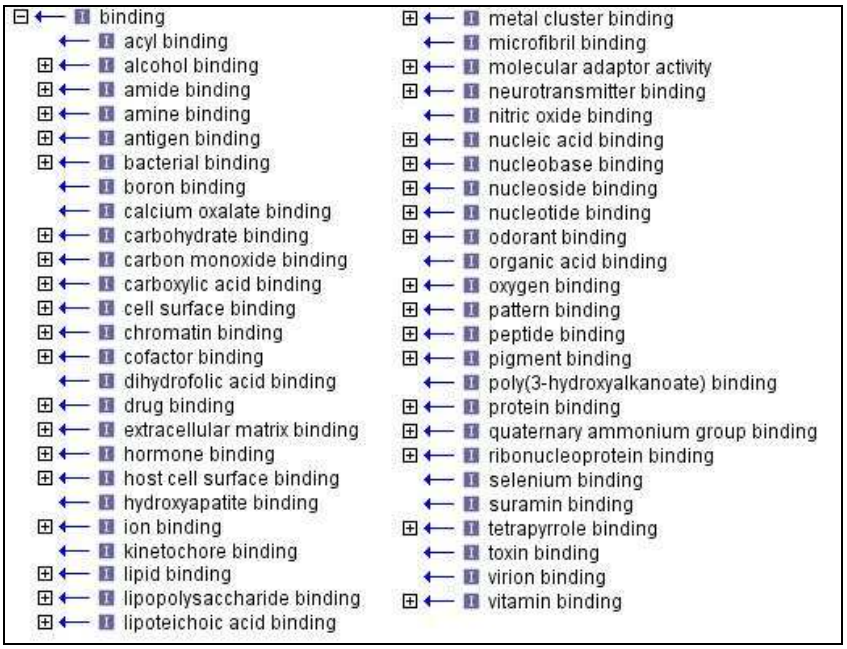


Figure 1. Subordinate classes of GO´s Molecular Function `binding` array

Results of the Analysis of the `Binding` array

Identification: GO:0005488 - Binding

Definition: *The selective, non-covalent, often stoichiometric* [1], i*nteraction of a molecule with one or more specific sites on another molecule*).

[1] The quantitative relation of the products and reactants of a chemical reaction in the proportion they appear in the chemical equation which describes the reaction [Smith et al., 2000].

Narrow Synonym: ligand

Classification[2] criteria observed:

- Chemical elements (ex: `boron binding`), organical compounds (ex: `lipid binding`), non organical compounds (ex: `nitric oxide binding`), kind of ion (ex: `ion binding`), organical radicals (ex: `acyl binding`), clusters of atoms (ex: `metal cluster binding`), role of molecules (ex: `antigen binding`), cellular locations (ex: `cell surface binding`).

In order to obtain the classificatory principles, besides analyzing the ontology hierarchy and terms definition, it was necessary to refer to specialized literature, in order to understand the nature of certain terms. For example, in the case of the term `Acyl binding` (Interacting selectively and non-covalently with an acyl group, any group formally derived by removal of the hydroxyl group from the acid function of a carboxylic acid), we came to the conclusion that "acyl" refers to a group of atoms (or radical), due to the fact that GOs[3] term definition refers to the term `acyl group`, which can be understood as a group of atoms or radicals (see definition of `group` and `acyl` in Oxford Dictionary of Biochemistry and Molecular Biology) [Smith et al., 2000] [4].

It is worth remembering that the analysis of GOs hierarchical structure, if carried out by a domain expert, or if applied more thoroughly and deeply in the ontologies hierarchies, could bring richer results, possibly with a wider range of observed problems.

Problems observed:

- **Violation of the Canon of exclusivity**

  Example:

  - **`norepinephrine binding`**: `Interacting selectively and non-covalently with norepinephrine, (3,4-dihydroxyphenyl-2-aminoethanol), a hormone secreted by the adrenal medulla and a neurotransmitter in the sympathetic peripheral nervous system and in some tracts of the CNS.`
  - This class is subordinate both to `alcohol binding` (binding/alcohol binding/norepinephrine binding) and to `amine binding` (`binding/amine binding/` norepinephrine binding).

- **Violation of the Canon of modulation**

  Example:

  - **`nitric oxide binding`**: `Interacting selectively and non-covalently with nitric oxide (NO).`

[2] For each organization principle observed, we have put in parenthesis, and italics, a term exemplifying those principles.
[3] Interacting selectively and non-covalently with an acyl group, any group formally derived by removal of the hydroxyl group from the acid function of a carboxylic acid.
[4] This dictionary appears as bibliographic reference within comments in GOs terms.

- Nitric oxide is a drug, according to the definition of the term `drug binding` (Interacting selectively and non-covalently with a drug, any naturally occurring or synthetic substance, other than a nutrient, that, when administered or applied to an organism, affects the structure or functioning of the organism; in particular, any such substance used in the diagnosis, prevention, or treatment of disease) and literature [Gerlach and Falke, 1995]. Therefore, according to the canon of modulation, the term `nitric oxide binding` should be subordinated to drug `binding`.

- **Violation of the Canon of subordinate classes**

  Example:

  - A `trisaccharide` (Interacting selectively and non-covalently with any trisaccharide. Trisaccharides are sugars composed of three monosaccharide units) is an `oligosaccharide` (Interacting selectively and non-covalently with any oligosaccharide, a molecule with between two and (about) 20 monosaccharide residues connected by glycosidic linkages), therefore its subordination to the class `sugar binding` (Interacting selectively and non-covalently with any mono-, di- or trisaccharide carbohydrate) violates the canon of subordinate classes, i.e., `trisaccharide binding` should be subordinated to the class `oligosaccharide binding`.

When tabulating the problems found, we highlight the importance of the canons of exclusivity, subordinate classes and modulation, as having the greater number of violation occurrences (7), followed by the canon of differentiation (5). There were not found evidences of violation of the canon of concomitance (although it is worth remembering that the analysis conducted did not cover thoroughly the complete deepness of GO´s hierarchies).

Table 1. Total occurrences found on analyzing GO first level classes

Canon	Total violations
Canon of Differentiation	5
Canon of Concomitance	0
Canon of Exclusivity	7
Canon of Modulation	7
Canon of Subordinate Classes	7

The analysis of first level hierarchies of GO´s Molecular Function branch shows a diversity of problems, which are materialized in a variety of non uniform classificatory principles observed, which seems to indicate a lack of adoption of well defined classificatory principles, gap which could be fulfilled by the adoption of Ranganathans' canons. In particular, we have observed the violation of Ranganathans' Exclusivity Canon, which points to the relevance of understanding existing perspectives to think

about the nature of the domains concepts. In contrast, we have not observed[5] the violation of the Canon of Concomitance. This finding could be evidence that some classificatory principles are more intuitively assimilated than others, but could be as well due to the characteristics of the domain, or yet, due to the deepness of the analysis conducted.

5. Conclusion

Ontology construction, although a maturing research field, still faces many challenges, especially in domains with a rich variety of complex concepts, and whose knowledge advances dynamically. In Biomedicine, for example, the need to organize concepts in a systematic way has to cope with the pragmatic nature of its community, with no deep knowledge of Ontology related disciplines such as Logics and Philosophy, but with urge to improve their ontologies.

One of the challenges of ontology construction is the creation of classificatory structures, or the backbone taxonomy, with its subclasses organized in a systematic way. The challenge presents itself not only due to the complex nature of the domains, but also due to the interdisciplinary nature of ontology building, which demands knowledge of experts in knowledge organization, such as Computer Scientists and Information Scientists, but especially, end users who detain the knowledge of the domain and the intended meaning of concepts contained in their ontologies. Considering that it is important to provide the grounds for an effective dialogue between people with different backgrounds and, at the same time, to provide principles that can rapidly and easily be assimilated and adopted on ontology structuring, it is important to overview existing and previously successfully adopted initiatives. In this sense, Information Science can provide relevant contribution, as it has mature hands on experience of information organization for more than fifty years, classifying subjects from many different areas.

Ranganathans' canons, some of those (but not all of them) were presented in this paper, can bring an important contribution to ontology structuring as it provides a comprehensive set of principles, that can be easily assimilated, and that provide effective guidance to avoid many of the problems found in ontologies structures, as we could observe by analyzing Gene Ontology. Although Ranaganathans' canons have been originally proposed long ago, they are a mature set of guidance that have been used successfully by more than fifty years, and are still current and relevant nowadays, as we have presented on our application scenario.

References

Berlin, B., Breedlove, D. E., & Raven, P. H. (1973) "General principles of classification and nomenclature in folk biology", American Anthropologist, n.75, p.214-242.

Bodenreider O, Smith B, Kumar A, Burgun A. (2004) "Investigating subsumption in DL-based terminologies: a case study in SNOMED CT", In: Hahn U, editor. KR-MED 2004; 2004. Whistler, Canada: AMIA; p. 12–20.

[5] It is worth noting that with the help of a biologist, with deep knowledge of the subject, other violations could have been found.

Campos, M. L. A. ; Gomes, H. E. (2008) “Taxonomia e Classificação: princípios de categorização”. Datagramazero (Rio de Janeiro), v. 9, n. 1.

Dahlberg, I. (1978b) “A referent-oriented, analytical concept theory of Interconcept”, International Classification, v. 5, n. 3, p. 122-151.

Dahlberg, I. (1983) “Conceptual compatibility of ordering systems”, International Classification, v. 10, n. 2, p.5-8.

Dahlberg, I. (1978a) “Teoria do conceito”, Ciência da Informação, v. 7, n. 2, p. 101-107.

Ekholm A. (2002) “Principles for classification of properties of construction objects”, Distributing Knowledge in Building - CIB W78 Conference.

Gene Ontology Consortium. (2001) “Creating the gene ontology resource: design and implementation”, Genome Research, v.11, n.8, p. 1425-1433.

Gerlach, H., Falke, K.J. (1995) “The therapeutic role of nitric oxide in adult respiratory distress syndrome”, Current Anaesthesia & Critical Care, v. 6, n. 1, p. 10-16.

Gomes, H.E., Motta, D.F., Campos, M.L. (2006) “Revisitando Ranganathan : a classificação na rede”, www.conexaorio.com/biti/revisitando.htm, Janeiro de 2011.

Guarino, N.; Welty, C., (2004) “An Overview of OntoClean”, In: S. Staab, R. Studer (eds.), Handbook on Ontologies, Springer Verlag, p. 151-159.

Guarino, N.; Welty, C., (2002a) “Evaluating Ontological Decisions with OntoClean”, Communications of the ACM, 45(2), p.61-65, 2002a.

Guarino, N.; Welty, C., (2002b) “Identity and Subsumption”, In: R. Green, C. A. Bean, S. Hyon Myaeng (eds.), The Semantics of Relationships: An Interdisciplinary Perspective, Kluwer, p.111-126.

Guizzardi, G., (2005) “Ontological Foundations for Structural Conceptual Models”, Doctorate Thesis in Computer Science, University of Twente, Enschede, Holand, ISBN 90-75176-81-3, 416 p.

Guizzardi, G.; Wagner, G.; Sinderen, M. (2004) “A Formal Theory of Conceptual Modeling Universals”, In: Proceedings Of The Workshop On Philosophy And Informatics (WSPI), Cologne, Germany.

ISO DIS 12006-2 - International Standards Organization. (1999) – “Organization of information about construction works — Part 2: Framework for classification of information”. Swiss.

Novellino, M.S.F. (1996) “Instrumentos e metodologia de representação da informação”, Informação & Informação., Londrina, v.1, n.2, p.37-45.

Ranganathan, S. R. (1967a), Prolegomena to Library Classification, New York : Asia Publishing House.

Ranganathan, S. R. (1967b) “Hidden roots of classification”, Information Storage and Retrieval, v. 3, n.4, p. 399-410.

Ranganathan, S.R. (1951), Philosophy of library classification. New Delhi: Ejnar Munksgaard.

Ranganathan, S.R. (1963), Colon Classification. Bombay: Asia Publishing House.

Smith, A.D.; Datta, S.P.; Smith, G.H., Campbell, P.N.; Bentley, R.; Mckenzie, H.A.; Bender, D.A.; Carozzi, A.J., Goodwin, T.W., Parish, J.H.; Stanford, S.C. (2000), Oxford Dictionary of Biochemistry and Molecular Biology, Oxford University Press Inc., New York.

Smith, B. (2005), "The Logic of Biological Classification and the Foundations of Biomedical Ontology", In: Westerstahl, D. (ed.) Invited Papers from the 10th International Conference on Logic, Methodology and Philosophy of Science.

Smith, B.; Kumar, A. (2004) "On Controlled Vocabularies in Bioinformatics: A Case Study in the Gene Ontology", BIOSILICO: Drug Discovery Today, v.2, p. 246–252.

Smith, B.; Williams, J.; Schulze-Kremer, S. (2003), "The ontology of gene ontology", In: AMIA Annual Symposium Proceedings, p. 609-13.

Yu, A.C. (2006), "Methods in biomedical ontology", Journal of Biomedical Informatics, n.39, p.252–66.

Realist representation of the medical practice: an ontological and epistemological analysis

André Q. Andrade[1], Maurício B. Almeida[1]

[1]Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6.627 – CEP 31270-901 - Belo Horizonte - MG – Brazil

andradeaq@ufmg.br, mba@eci.ufmg.br

Abstract. *Realist ontologies organize knowledge by strict adherence to philosophical principles, ensuring robustness and coherence. According to those principles, only entities empirically verifiable can be represented. Our study aimed to analyze medical records to evaluate which kinds of entities should be represented for physicians. We classified the entities and found several entities that cannot be represented in realist ontologies. After due analysis, results suggest that a categorization that distinguishes reality from medical knowledge about reality and observations under both of them are useful to describe entities present in medical records.*

1. Introduction

Information structuring in electronic health records (EHR) is essential for the development of health applications, due to its ability to exchange information between different applications and healthcare professionals. Structured records are amenable to use in several situations, such as: a) scientific discoveries; b) use of recorded data by other professionals; c) healthcare facility management and quality control; d) prevention of epidemics and health policy development.

System interoperability (the capacity of communication between systems without human intervention) requires shared semantics of terms used in both systems. Recently, the use of ontologies for semantic representation is being studied in several domains, like the biomedicine [Rubin et al. 2008]. Particularly, the development and wide adoption of the realist stance for ontology creation allows for an explicit, stable and language independent vocabulary definition, which promotes communication without ambiguities [Smith and Ceusters 2010].

Even though such methodology aims to describe scientific knowledge, such as gene and protein biological functions [Hill et al. 2008], it actually limits the representation of natural language terms that have no direct referent in the world. This "non-ontological" terminology is important to clinical records [Stenzhorn et al. 2008]. For instance, a clinician may use the term "hepatitis" to refer to a real hepatitis, but can also refer to a clinical suspicion (that the patient may have the disease), or to a preventive action (like vaccine prescription). In this paper we generically refer to those terms as epistemological [Bodenreider et al. 2004].

Besides, Schulz and colleagues [2009] argue that the attempt to code probabilistic and default knowledge using ontologies is likely to create incorrect models. In fact, the realist approach seems incapable to represent statements that are not universally true, such as "suspected fever", "past history of fever" and "no fever".

This research aims to evaluate real medical records, to analyze which entities are amenable to representation by the so-called realist ontologies, as defined by [Smith 2006]. The description of medical record entities by ontological and epistemological principles, part of an ongoing research project, is being used to create a set of procedures that will guide the analysis and create a generic framework that will improve understanding of medical systems specificities.

This paper is structured as follows. In section 2 we describe the advantages and limitations of realist ontologies for medical knowledge representation. In section 3, we present a critical evaluation regarding the relationship of formal ontologies and clinical reasoning. In section 4 we present the methodology used, aimed at identifying information contained in real medical records. In section 5 we present results, in section 6 we discuss the results and in section 7 we present our final remarks.

2. Ontologies

Ontologies are being used in large scale in varied domains like architecture, geography, [Bittner 2010], medicine and biology [Bittner and Donnelly 2007], whether as support for legacy classification systems, or as way of adequately representing a domain. In the following sections we describe applications in biomedical ontologies (section 2.1), as well foundational principles of realist ontologies, widely used in biomedicine (section 2.2).

2.1. Biomedical ontologies

Ontologies have been successfully used in the biology and medical domain around the world. Several initiatives were gathered in the *Open Biomedical Ontologies Foundry* (OBO), a repository of accessible, interoperable ontologies, described in uniform syntax and unequivocal identification [Smith et al. 2007]. Considering the OBO group, some ontologies are worth mentioning, due to innovation and intense use in scientific research. Among them, the *Gene Ontology*, an ontology that describes basic characteristics of genes; the *Foundational Model of Anatomy*[1], which describes the prototypical human anatomy; the *Cell-Type Ontology*, which describes cell types from some living; the *Protein Ontology*[2], which describes the relationship between proteins and classes that represent protein evolution; and the *Chemical Entities of Biological Interest*[3].

Besides these big foundational ontologies, several others are still under evaluation and available at OBO, such as the *Disease Ontology*[4] [Cowell and Smith 2010], the *Ontology for Biological Investigations* [Brinkman et al. 2010], the *Ontology for General Medical Science*[5] [Scheuermann et al. 2009].

[1] http://fma.biostr.washington.edu/
[2] http://pir.georgetown.edu/pro/
[3] http://www.ebi.ac.uk/chebi
[4] http://do-wiki.nubic.northwestern.edu/index.php/Main_Page
[5] http://code.google.com/p/ogms/

2.2. Realistic ontologies

The term "realism" in Philosophy is widely used and controversial [Miller 2010]. We have to emphasize that realism, while philosophical discipline, can disclose different flavours. Indeed, there are issues under unending debate among people which declare themselves as being realists. Defining universals, a main tenet of realism, is an example of issue on which there is no agreement [MacLeod 2005]. In this paper, we take the "ontological realism" as a methodology for ontology development – said "realist ontologies" – based on principles of the philosophical realism. It is a methodology widely used in biomedicine [Baker et al. 1999][Grenon et al. 2004] grounded at the following generic tenets: [Munn and Smith 2008]: i) there is a real world; ii) the reality in which we live in is part of this world; iii) we are capable of knowing the world and reality, even if just in a approximate way.

One of the assumptions of the ontological realism is the theory of universals, which states that in reality there are particular and universal entities. Particulars are entities described by the observation of the real world, e.g. a clinic or a laboratory. Universals represent that which is common to every correspondent particular - e.g. the characteristic of having a head that is common to every human being – which is invariant in reality [Smith 2004][Smith 2006]. Since ontological realism is based on reality and proposes that the best way to describe it is through science, universals are those entities chosen to be used in the formulation of scientific theories.

According to the ontological realism, the unrestricted creation of classes to represent every possible entity leads to inconsistencies. Classes are human creations – e.g. every human being that is a man and likes swimming – and may be interpreted in different ways [Munn and Smith 2008]. To avoid that, the realist methodology restricts the possible classes to those defined by the scientific community. However, the precise distinction between universals and classes is not always trivial. While universals are grouped by what they are, classes are grouped by how they are [Smith e Ceusters, 2010].

The realist methodology uses a upper-level ontology to organize universals with a top-down approach. Examples of upper-level ontologies are the BFO [Grenon et al. 2004], DOLCE, the SUMO, among others. In the BFO, adopted in the ontological realism stance, we can find structuring divisions made by generic universals called continuants and occurrents. This division is based on the notion of SNAP and SPAN [Grenon et al. 2004]. SPAN entities, called occurrent or perdurants, are universals that posses a determined beginning and end, and encompass process (e.g. "the life of an organism") and spatiotemporal regions (e.g. "the eighties"). SNAP entities, also called continuants or endurants, are universals for particular that maintain their identities through time (e.g. a "human being"). Continuants may be dependent (e.g. "the color of an object"), independent (e.g. "a table") or spatial regions (e.g. a "point"). To explain the different treatments for high-level entities in other ontologies abovementioned is beyond the goals of the present paper.

The use of the same upper-level ontology as starting point to create domain ontologies increases the chance that its universals are compatible and, therefore, the chance that they are amenable to integration.

3. The limitations of realist ontologies for representing medical practice

The extension of realist biomedical ontologies to the medical practice is an alternative for medical information organization. However, considering institutions of most countries, medical documentation is usually made of barely structured documents, sometimes even handwritten, containing heterogeneous information. Even so, the medical record is an essential work tool for the clinician. The record is used for medico-legal reasons, as a tool to support care plan creation and as a support to find information required for clinical decision-making.

The realist ontology approach has been the target of many criticisms, which usually argue against the proposal of universals as a sine qua non condition to the creation of good ontologies [Merrill 2010a][Merrill 2010b] [Rector 2010] [Cimino 1998][Cimino 2006] [Dumontier and Hoehndorf 2010]. Such approaches for biomedical ontologies emphasizes the importance of language, communication and medical reasoning and puts under suspicion the obligation in considering [Merrill 2010a]. In many cases, such approaches have been labeled as "epistemological".

Conceptual approaches, a variant of idealism, are closer to medical everyday language, since they use terms not referenced in reality which are commonly present in new and yet not fully comprehended clinical situations. In the medical practice, diagnoses are usually presumptive and based on incomplete data, making it difficult to identify a particular and the corresponding universal. In fact, statements in such context are constantly revised and do represent truths, but the physicians grounded opinion.

Realist-oriented researchers argue that the creation of ontologies around concepts is based on language and, therefore, is subject to ambiguities and differences of understanding and interpretation by different individuals [Smith 2006]. These researchers consider that ontologies are artifacts made for use by computers and that any natural language-derived ambiguity harms interoperability efforts. This is particularly important in natural sciences representation such as biology in which, despite the enormous volume of data, there is consistency in observation by different institutions. Also in medicine, anatomical and physiological statements are consensual when attributed to universals.

This is made clear by comparing the statements "AIDS is spreading quickly through Asia" and "AIDS is caused by the HIV". The term AIDS in the former is a class, while it correspond to a universal in the latter. Classes are arbitrary sets and can result in representation that cannot be understood and interpreted. By restricting the ontological commitment to reality as described by science, the ontological realism promotes consensus.

Another relevant aspect to be considered is the distinction between ontology and epistemology. Epistemology is the study of how cognoscent beings come to the truth about some event in reality. The difference between the terms can be shown by evaluating how entities are defined in ontology and in epistemology. Ontology is about an object, process, event, whole, part, determination, dependence, composition, etc. Epistemological statements are about the way we know things and is about belief, truth, probability, confirmation, knowledge and its variations [Poli 2010]. While ontology is a theory of things, epistemology is a theory of knowledge.

The interdependence between the existence of an entity and the knowing about it frequently blurs the distinction between ontology and epistemology. Bodenreider and colleagues classify epistemological terms usually identified in biomedical terminologies in four categories [Bodenreider et al. 2004]:

- Terms containing classification criteria: terms that do not represent universals, but that intend to convey information. For example, the distinction between "febrile seizure" and "afebrile seizure" is not a distinction between characteristics of the seizure itself, but conveys information about probable cause and prognosis.
- Terms reflecting detectability, modality, uncertainty, and vagueness: since complete understanding of a clinical situation is very difficult, physician usually express this incomplete knowledge of the patient condition by modal and approximate statements. E.g. "possible cancer", "probable cancer", "unspecified chest pain".
- Terms created in order to obtain a complete partition of the domain: contain terms that intend to encompass entities not described by other classes. E.g. "Other" and "Pneumonia not otherwise specified".
- Issues related to normality and to fiat boundaries: terms that intend to convey instructions about how the information should be interpreted, not about the entity itself. E.g. "normal height", "enlarged liver". It is important to point out that part of the medical knowledge is based on historical events which had an almost arbitrary definition of normality [Vickers et al. 2008].

The fact that clinical observations are necessary is not opposed to the realist methodology: information about opinions are fundamentally different from information about objects [Munn and Smith 2008] and both have a place in an descriptive ontology[6]. However, the medical practice requires the recording of information of both natures, named here ontological and epistemological, including impressions, plans, suggestions, etc. We intend to pursue this issue while searching for a complementary approach that helps in understanding the medical reality.

4. Methodology

This ongoing research objective is to evaluate the representation of health information in real medical records, through the use of realist ontologies. We intend to determine its limitations and propose new ways of representing non-ontological information. For example, administrative data, which at first had no counterpart in realist reference ontologies, has to be represented through the creation of other ontologies for dealing with such entities, like the Information Artifact Ontology [IAO 2011]. The methodology is composed by the following steps:

1. Record creation based on real clinical cases: The analysis must consider the way health professionals record medical events. We studied two complete records, created by two Internal Medicine specialists, based on common presentations of real patients. No identification data was recorded, such as name, age,

[6] "concerns the collection of such prima facie information on types of items either in some specific domain of analysis or in general"[Poli 2010, pg.2]

geographical location, health facility, dates and identification and contact numbers, according to recommendations by [Meystre et al. 2010].

2. Transcription of records for information identification: In order to identify information unities, a domain expert transcribed the records in sentential fragments. The domain expert was asked to identify the reason for recording those entities and the information that is being conveyed by the representation. The transcription used the principles of logic and controlled languages described [Fuchs et al. 2005][Fuchs et al. 1999] , which allowed clear identification of entities recorded in natural language, outside the particular context in which the event took place [Vickers et al. 2008]. Since the objective of this paper is to analyze the content of the text, syntactical and markup aspects pertinent to automatic processing are omitted. We hereafter call those information unities as entities, despite their physical existence.
3. Analysis and classification of the record's information items, according to ontological realism guidelines: The information entities were analyzed according to the tenets of the ontological realism [Grenon et al. 2004] , to verify if they were suitable to ontological representation. This analysis was guided by pre-established criteria aimed to classifying the entity in some upper BFO class. Some examples can be found in table 1. Each entity was tested against the set criteria, respecting the BFO class hierarchy. E.g. the first test separates entities in continuants and occurrent; after this distinction, specific criteria are used for each class. The entities that don't belong to any BFO class are analyzed according to realist principles and their use in everyday medical practice. We selected some cases for further discussion, presented in section 6.

Table 1. Distinction between continuants (EMT) and occurrents (ECT)

Distinction (I)	Entities that maintain their identity through time (EMT)	Entities that change through time (ECT)
Characteristics	a) The entity exists completely in any given period of time in which it is present b) The entity has no temporal parts.	a) The entity unfolds through a period of time.

5. Results

The records analyzed represent outpatient visits. The first one describes the consultation of a patient with an unexplained chest pain and the second a post-discharge consultation due to dyspnea. The records make use of routine record organization, such as "Complaint", "History", "Physical examination", etc. In table 2 we present a small extract of one of the documents. Partial results can be seen below in Table 3:

Table 2. Extract of an outpatient record of a fictitious patient

QP: Chest pain and abdominal pain. HMA: Six months ago, the patient felt severe precordial pain in addition to nausea and dyspnea. She attempted medical care in the Hospital X, where received isordil + AAS 300mg. Enzimes: CKT 262 CKMB 30. She was not aware of previous pathologies. It was prescribed: Captopril, HCTZ e AAS. Last month, the patient felt severe pain again and sought for medical care in a different place. Then, it

was prescribed: Losartan, AAS, Sinvastatina e Nebilet. She sought for medical care in other occasions because of the precordial pain. In addition to the medicine mentioned, she uses Metoprolol - 100 mg 12/12 h. She reports diffuse and intermittent abdominal pain, which becomes worse in case of stress. It is not related with bowel movement alterations. She also reports rare burning epigastric pain that improves with water drinking.

Table 3. Example of mapped and non-mapped entities to realist ontologies

1- Aspects that represent entities IN REALITY (some examples)	
Continuant	**Occurrent**
-Chest pain -Abdominal pain, Precordial pain, Epigastric pain -Nausea, Dyspnea -Enzyme -Captopril, Losartan	-Were prescribed -Makes use -Bowel movements -Moment of first occurrence of pain (six months) -Moment of re-incidence of pain (one month ago)
2- Aspects that represent useful constructs for medical practice NOT empirically verifiable	
-**Severe** (precordial) **heavy pressure** (pain)	-**Diffuse and intermittent** (abdominal pain) -**Rare burning** (epigastric pain)
3- Aspects that represent observations ABOUT reality (not reality itself)	
-CKT 262 -CKMB 30 -Left ventricle ejection fraction: 68%	
4- Aspects that represent observations ABOUT the physician understanding of the clinical situation (not about reality)	
-Previous consultations and prescriptions -Not related to bowel movement alterations	-Previous diseases - (Diffuse and intermittent abdominal pain) **that worsens with stress** - (Rare burning epigastric pain) **that improves with water drinking**

6. Discussion

The medical record is a complex document used for several purposes in healthcare processes. According to the Brazilian Medical Council, it is "a single document made of a set of recorded information, signs and images, created after (events) about the patient health and care provided, of a legal, private and scientific character, that allows communication between the multi-professional team and continuity of care provided to the individual" [Conselho Federal de Medicina 2002, art 1º]. To live up to those expectations, the professional uses the flexibility of natural language expressions to represent the clinical situation, his clinical reasoning process and the relevant context of the health event.

In our research, we drew terms from records trying to fit them to constraints imposed by realist ontologies. Then, we created two main sets: in the first one, we included the entities that could be represented in realist ontologies; the second one gathers entities that can not be represented in realist ontologies as we defined them in the context of this paper (vide section 1) Terms that can be used in realist ontologies are presented as the first group of table 3. Arguably, realism has been shown capable of representing diseases, disorders [Scheuermann et al. 2009] and symptoms [Smith et al. 2009], as evidenced by the *Ontology for General Medical Science* (OGMS). The existence of diseases – defined as a "disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism" [Scheuermann et al.

2009, pg.3] is well known by medical science, and its representation is robust and homogenous. Likewise, symptoms can be seen as body characteristics that a patient experiences. In this case, we represent the body alteration considering its scientific description. On the other hand, the diagnosis itself is not a patient attribute, but rather "a conclusion of an interpretive process that has as input a clinical picture of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a disease of such and such a type." [Scheuermann et al. 2009, pg.5])

We observed that the realist methodology in incapable of defining symptoms qualities. In order to evaluate a patient, each symptom must be described according to its seven characteristics [Bickley and Szilagyi 2009]: Location; Quality; Quantity or severity; Timing; Setting in which it occurs; Remitting or exacerbating factors; Associated manifestations. These characteristics can be classified in three groups, according to their relation with realist ontologies.

Formal ontologies are capable of precise representation of symptom location and temporality, through the description of body structures – organs and systems – or spatiotemporal regions. This first group can be described by upper level classes of the BFO, as continuants – independent continuants and spatiotemporal regions – and occurrents – the temporal region occupied by the symptom and, eventually, the symptom itself. For example, "chest pain" and nausea". As stated in the Methodology section, this analysis was based on the BFO, but different upper ontologies may suggest different approaches. This is markedly true in the case of qualities [Masolo and Borgo 2005], defined by the BFO as "a specifically dependent continuant that is exhibited if it inheres in an entity or entities at all (a categorical property)" [Basic Formal Ontology].

The second group of table 3, containing the characteristics of quality and quantity/severity, describes attributes of the symptoms, its temporal evolution, qualities, dispositions, functions and roles. The qualities refer to symptom types as described by scientific knowledge of common clinical presentation of diseases. There are regional and national variations of such typology, but classical symptom description is fairly constant. For instance, the term "crushing pain" is commonly interpreted as a cardiac originated pain. The quality "crushing" of the precordial pain has no direct and unequivocal relation with the subjacent disorder, but the history of pain is similar in patients with the same kind of disorder. In this case, we argue that the term is not a realist universal, but can be described by concepts in a coherent fashion. The same criteria applies to severity ("Severe pain"), which shows the same linguistic ambiguity (how much is severe?). These terms must be described by non-ontological artifacts to avoid reasoning and classification errors, since the distinction between types of pain cannot be empirically verifiable – e.g. distinguishing a crushing from a heavy pressure pain. Another example would be "diffuse abdominal pain", which should not be treated as a single ontological entity.

In the fourth group of table 3, we find the aspects referring to the situation in which the symptom was experienced by the patient, according to the medical record description. The setting of occurrence describes the state of affairs at the moment when the symptom was perceived, what the patient was doing, climate and environment conditions, events that preceded the symptom, etc. Remitting or exacerbating factors describe entities that, according to the patient's or physician's interpretation, changed the natural course of the symptom. This interpretation may be motivated by previous knowledge (e.g. causes of

chest pain may be distinguished by their relation with physical exercise), temporal coincidence or unjustified beliefs. Finally, associated manifestations may be represented by any other symptom, or the absence of symptoms, as long as they aid medical reasoning and the definition of a diagnosis. The representation of this group through realist ontologies is mostly ambiguous. In the cited examples, the occurrent "drinking water" and the "epigastric pain" intensity decrease are temporally related, but the causality cannot be empirically determined. Rather, they reflect the understanding of the situation, so that there is a belief that both entities are causally related.

Besides symptoms, several other entities were found, such as medications, laboratory test results, physical examination findings, among others. Entities like life signs measurements and lab test results do not directly refer to patient qualities, but to observations about those qualities. For example, the CKMB (creatine phosphokinase MB) refers to the enzyme blood concentration at the exact moment of blood sample collection. It is, therefore, empirically verifiable. However, the value of the measurement is arbitrarily determined (in this case, unity per liter) and does not refer to the existence of the enzyme in the real world. Besides, this information is not analyzed using logic operations, but used in a sequence of pre-established thinking rules, according to clinical training. In this clinical case, the value 30 U/L is just above the normal value (26 U/L) and, therefore, leads the physician to question the hypotheses of myocardial infarction, suggested by the initial presentation of chest pain. The presence of continua in the real world requires fiat delimitations, which are justified by pragmatic reasons. [Schulz and Johansson 2007]. We argue that this information should be distinguished from direct referents, since it refers to a representation of an observation about reality. Moreover, it will be interpreted according to reasoning structure, not according to the structure of reality itself.

Several solutions to this problem can be found. Despite conceptualism shortcomings, the restrict use of concepts to represent epistemological information expands the scope of those representations. To improve the meaning of concept and avoid misuse of the term, we will use the definition put forward by [Klein and Smith 2010, pg.722]: "concept should be used exclusively to refer (1) to the meaning of a corresponding general term, this meaning being (2) unique and (3) agreed upon by responsible persons in the given disciplinary field."

The use of concepts to represent clinical information, though subject to inconsistencies, is closer to language, since it represents term meaning and does not denote an entity (universal or particular). The definition of each term can be made through formal languages or natural language description, depending on the heterogeneity of interpretations given to a term: e.g. concepts such as "up" and "down" are intuitive and interpreted in a constant way, while the term "AIDS" requires precise explanation. The relation between terms should be done through semantic relations "broader_than" and "narrower_than", considering the term meaning. Additionally, we can consider the relation "related_to", as proposed in the W3C *Simple Knowledge Organization System* (SKOS) standard.

While the proposed typology encompasses real and epistemological entities, it still needs improvement. The description of knowledge, thought as an attribute of the cognoscent being, does not describe a real entity, but says something about it. For instance, it is false to represent a "canceled surgery" as a "surgery", for it never

happened. A solution is to represent the “canceled surgery” as an information artifact, a plan that is about the “surgery”. In this case, the surgery will never come to be, but the plan existed through a defined and verified temporal region [Schulz et al. 2010]. These and other hypothetical entities, like instructions (“in case of recurrence, do X”) and goals (“the patient should try to lose at least 2 kg with this diet within 2 months”) can be placed in more than one category – it may be seen as a real information entity ABOUT another not yet instantiated real entity, or as a mental model simulation on the part of the physician, stating an algorithmic behavior in the form IF X THEN Y. These cases make evident that further effort in refining the model must be done.

7. Final remarks

The proposed categorization suggest that understanding reality representation in four levels – reality itself, the perception of reality by the being, and the recording of reality [Smith et al. 2006] – makes clear the connection between representation methodologies. This connection will be important for proper computerizing of medical records.

During this research project, we intend to expand this categorization in a framework connecting ontological and non-ontological entities that promotes representation of entities required in medical practice without compromising interoperability and automatic inferences. We intend to explore information models such as the HL7 v3 and the OpenEHR, since they offer a great opportunity to understand the relation between concepts and ontological entities.

8. References

Baker, P. G., Goble, C. A., Bechhofer, S., Paton, N. W., Stevens, R. and Brass, A. (1999) "An ontology for bioinformatics applications". Bioinformatics 15 (6):510-520

Basic Formal Ontology "Basic Formal Ontology v1.1.1". http://www.ifomis.org/bfo/owl. Accessed June 18th 2011

Bickley, L. S. and Szilagyi, P. G. (2009) Bates' Guide to Physical Examination and History Taking. 10th Edition edn. Lippincott Williams & Wilkins,

Bittner, T. (2010) "On the integration of regional classification systems for the National Map". Cartographica: The International Journal for Geographic Information and Geovisualization 45 (2):127 - 139

Bittner, T. and Donnelly, M. (2007) "Logical properties of foundational relations in bio-ontologies". Artif Intell Med 39 (3):197-216. doi:10.1016/j.artmed.2006.12.005

Bodenreider, O., Smith, B. and Burgun, A. The Ontology-Epistemology Divide:A Case Study in Medical Terminology. In: A. Varzi, L. Vieu (eds) 3rd Conference on Formal Ontology in Information Systems, Turin, 2004.

Brinkman, R. R. et al. (2010) "Modeling biomedical experimental processes with OBI". J Biomed Semantics 1 Suppl 1 (22):S7

Cimino, J. J. (1998) "Desiderata for controlled medical vocabularies in the twenty-first century". Methods Inf Med 37 (4-5):394-403

Cimino, J. J. (2006) "In defense of the Desiderata". J Biomed Inform 39 (3):299-306

Conselho Federal de Medicina (2002) Resolução 1638/2002.

Cowell, L. G. and Smith, B. (2010) Infectious Disease Ontology In: V. Sintchenko (ed) Infectious Disease Informatics. Springer, pp 373-395

Dumontier, M. and Hoehndorf, R. (2010) "Realism for scientific ontologies". Paper presented at the Proceeding of the 2010 conference on Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), Toronto, Canada,

Fuchs, N. E., Hofler, S., Kaljurand, K., Rinaldi, F. and Schneider, G. (2005) "Attempto controlled english: A knowledge representation language readable by humans and machines". Reasoning Web 3564:213-250

Fuchs, N. E., Schwertel, U. and Torge, S. (1999) A Natural Language Front-End to Automatic Verification and Validation of Specifications. LMU München,

Grenon, P., Smith, B. and Goldberg, L. (2004) "Biodynamic ontology: applying BFO in the biomedical domain". From D M Pisanelli (ed), Ontologies in Medicine, Amsterdam: IOS Press, 2004, 20–38

Hill, D. P., Smith, B., McAndrews-Hill, M. S. and Blake, J. A. (2008) "Gene Ontology annotations: what they mean and where they come from". BMC Bioinformatics 9.

IAO (2011) "Information Artifact Ontology". http://code.google.com/p/information-artifact-ontology/. Accessed June 18th 2011

Klein, G. O. and Smith, B. (2010) "Concept Systems and Ontologies: Recommendations for Basic Terminology". Information and Media Technologies 5 (2):720-728

MacLeod, M. C. (2005) "Universals". http://www.iep.utm.edu/universa/. Accessed June 28th 2011

Masolo, C. and Borgo, S. (2005) "Qualities in formal ontology". Paper presented at the Foundational Aspects of Ontologies (FOnt 2005) Workshop at KI 2005, Koblenz, Germany,

Merrill, G. H. (2010a) "Ontological realism: Methodology or misdirection?" Applied Ontology 5:79-108

Merrill, G. H. (2010b) "Realism and reference ontologies:Considerations, reflections and problems". Applied Ontology 5:189–221

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. and Samore, M. H. (2010) "Automatic de-identification of textual documents in the electronic health record: a review of recent research". BMC Med Res Methodol 10:70

Miller, A. (2010) "Realism". http://plato.stanford.edu/entries/realism/. Accessed July 25th 2011

Munn, K. and Smith, B. (eds) (2008) Applied Ontology. An Introduction. Ontos Verlag, Frankfurt/Paris/Lancaster/New Brunswick

Poli, R. (2010) Ontology: The Categorial Stance. In: R. Poli, J. Seibt (eds) Theory and Applications of Ontology: Philosophical Perspectives. Springer, Berlin, pp 1-22

Rector, A. (2010) "Knowledge Driven Software and "Fractal Tailoring": Ontologies in development environments for clinical systems". Paper presented at the Proceeding

of the 2010 conference on Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), Toronto, Canada,

Rubin, D. L., Shah, N. H. and Noy, N. F. (2008) "Biomedical ontologies: a functional perspective". Brief Bioinform 9 (1):75-90. doi:10.1093/bib/bbm059

Scheuermann, R. H., Ceusters, W. and Smith, B. Toward an Ontological Treatment of Disease and Diagnosis. In: 2009 AMIA Summit on Translational Bioinformatics, San Francisco, CA, 2009. pp 116-120

Schulz, S. and Johansson, I. (2007) "Continua in biological systems". The Monist 90 (4)

Schulz, S., Schober, D., Daniel, C. and Jaulent, M. C. (2010) "Bridging the semantics gap between terminologies, ontologies, and information models". Stud Health Technol Inform 160 (Pt 2):1000-1004

Schulz, S., Stenzhorn, H., Boeker, M. and Smith, B. (2009) "Strengths and limitations of formal ontologies in the biomedical domain". RECIIS Rev Electron Comun Inf Inov Saude 3 (1):31-45

Smith, B. Beyond concepts: Ontology as reality representation. In: A. Varzi, L. Vieu (eds) FOIS 2004, Turin, , 4-6 November 2004. IOS Press, pp 73-84

Smith, B. (2006) "From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies". Journal of Biomedical Informatics 39 (3):288-298.

Smith, B. et al. (2007) "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". Nature Biotechnology 25 (11):1251-1255.

Smith, B. and Ceusters, W. (2010) "Ontological realism: A methodology for coordinated evolution of scientific ontologies". Applied Ontology 5:139–188

Smith, B., Ceusters, W., Goldberg, L. J. and Ohrbach, R. K. (2009) "Towards an Ontology of Pain and of Pain-Related Phenomena".

Smith, B., Kusnierczyk, W., Schober, D. and Ceusters, W. Towards a reference terminology for ontology research and development in the biomedical domain. In: KR-MED 2006, 2006. pp 57-66

Stenzhorn, H., Schulz, S., Boeker, M. and Smith, B. (2008) "Adapting Clinical Ontologies in Real-World Environments". J Univers Comput Sci 14 (22):3767-3780

Vickers, A. J., Basch, E. and Kattan, M. W. (2008) "Against diagnosis". Ann Intern Med 149 (3):200-203

Ontology Enrichment Based on the Mapping of Knowledge Resources for Data Privacy Management

Fernando M.B.M. Castilho[1], Roger L. Granada[1], Renata Vieira[1], Tomas Sander[2], Prasad Rao[2]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Ipiranga Av., 6681. FACIN. CEP 90169-900. Porto Alegre, Brazil.

[2]Hewlett-Packard (HP)
Ipiranga Av., 6681. Building 91B. CEP 91530-000. Porto Alegre, Brazil.

{fernando.castilho, roger.granada}@acad.pucrs.br,
renata.vieira@pucrs.br, tomas.sander@hp.com, prasad.rao@hp.com

Abstract. *This paper presents a mapping of enriched knowledge resources for data privacy management. An ontology, enriched through natural language techniques, is used for an integrated visualization for global inspection of heterogeneous data. The visualization helps stakeholders in exploring and maintaining a knowledge base for data privacy accountability. The integration of resources on the basis of concepts described in an enriched ontology is an aid to Knowledge Management (KM) in a dynamic domain, due to changes in laws and the corresponding system requirements.*

1. Introduction

The use of ontologies helps to achieve consensus on terms related to specialized domains. The mapping of heterogeneous resources from knowledge rich systems can help domain stakeholders in achieving their knowledge intensive related tasks. This paper is contextualized in the data privacy domain, especially considering the task of accountability. One of the main concerns in data privacy accountability is to avoid data misuse in collecting and handling Personal Identifiable Information (PII) [1]. To ensure that an organization needs robust mechanisms to implement its privacy policies.

Weitzner [2] defines: "Information accountability means that information usage should be transparent so it is possible to determine whether a use is appropriate under a given set of rules". One aspect of determining such usage is the identification of privacy risks related to sensitive information[1]. We discuss the integration of knowledge resources of a rule based tool that provides guidance and privacy assessment of a project that handles PII and identifies possible privacy risks. From now on we simply call it 'accountability tool'. Its resources comprise a questionnaire, a glossary of privacy terms, a set of encoded rules, company policies and a set of guidelines for developers.

The motivation of our work relies on the enrichment of an ontology, based on linguistic and knowledge resources in the privacy domain. We developed a visualization tool to integrate these resources. The mapping of knowledge sources and artifacts, based on an ontology model, can provide a better overview of the information handled by the

[1] "Sensitive information" as defined in: TCSEC - Department of defense trusted computer system evaluation criteria. Dept. of defense standard, Department of Defense, Dec 1985.

accountability tool, and thereby support various critical tasks to reduce oversights and errors in the management of privacy in company projects.

The domain stakeholders are privacy officers, Knowledge Base (KB) engineers, and project managers. Privacy officers are generally accountable for compliance with privacy regulation, and for creating, maintaining and checking the correctness of the underlying KB, as well as for evaluating impacts of changes in the body of laws and documents such as company privacy guidelines. They are in charge of transforming laws and regulations into specific company policies and guidelines. KB engineers are in charge of modeling legal constraints and requirements involving policies and laws, and writing and updating rules in the accountability tool. Finally, project managers are responsible for company projects and their alignment with organizational policies. As an example of benefits of the KM, richer information may help project managers to take information into account such as privacy lawsuits in progress, upcoming changes in laws and regulations etc. that are otherwise unlikely to be available to them.

To help stakeholders with their tasks we propose a mapping between various knowledge sources. The existing sources comprise privacy regulation documents, along with the KB of the rule based system mentioned earlier. Privacy documents are regulatory texts like acts, norms and guidelines for privacy assurance and safe, and accountable software development[2]. A domain ontology was developed, which is at the core of the mapping structure. It is enriched by automatically generated resources: a thesaurus and a list of Named Entities (NE) referring to normative regulations in the privacy domain. The idea is that the enriched ontology can serve to maintain the rule based system. Our work is then based on the definition of an automatically enriched conceptual structure, and on the mapping of knowledge sources to ontology concepts, aiming at the establishment of a KB management infrastructure.

This paper is organized as follows: Section 2 introduces related work, with an analysis of the privacy risk management problem and the use of ontologies in this area, and describes the contribution of our model for the representation of data privacy risks. Section 3 presents an overview of linguistic and semantic resources and their integrated visualization based on the enriched ontology. In Section 4 the overall process of integration of the resources is presented, along with the evaluation of the ontology enriching methods for Thesaurus and NE. Finally, in section 5 we present our concluding remarks.

2. Related work

The development process in Information Technology (IT) is one of the main areas on which privacy strategy needs to focus. Taking up to date privacy legislation into account is an important requirement for IT projects. The knowledge of weaknesses in projects with respect to privacy laws and guidelines, as well as their correct application in such projects, can help to correct inadequate procedures or prevent serious privacy incidents such as data breaches, and thus avoid lawsuits and the loss of consumer trust for a company [3].

[2] Asia-Pacific Economic Cooperation. 2004. APEC Privacy Framework. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector, Official Journal L No. 201, 31.07.2002. Microsoft Privacy Guidelines for Developing Software Products and Services, v. 3, 2009. The U.S. Children's Online Privacy Protection Act of 1998 (COPPA). The U.S. Computer Fraud and Abuse Act. U.S. Federal Trade Commission (2000). Financial Privacy Rule.

Our work aims at the identification of privacy risks, and the management of the domain KB that supports it. An ontology was modeled considering the movement of information across borders, and the actions performed on it to identify privacy risks, and to organize thinking and discussion in the privacy field, which is relevant to IT. This approach follows the need stated by Solove [4] who developed a taxonomy of privacy to describe concepts of information collection, processing, dissemination and invasion to capture violations of privacy. It focuses on actions and focuses on activities such as the collection, processing and dissemination of information, which remove it further from the direct control of the user.

A process to promote privacy assurance inside organizations and to establish proper privacy management to address legislative requirements, policy guidance and business standards is proposed in [5]. Knutson [6] presents some principles that organizations should follow to create privacy awareness. He points out that a privacy core team with technical and legal experts must define a privacy terminology to achieve a common understanding of the scope and meaning of rules. Another recommendation is to create guidelines to help developers to become independent from privacy experts with respect to basic tasks. Similar concerns for software design are endorsed within other works on privacy awareness [7][8]. In our work these requirements are carried out with the definition of an ontology enriched by integrated knowledge resources.

Recognizing concepts and instances in text in order to support ontology maintenance and semantically represent the meaning of sentences is a task explored in [9]. One step towards a better control of the development process from a privacy perspective is to have a proper representation of the relevant rules that have already been formulated for handling PII. These rules are mostly described in laws, policies and other normative sources, such as implementation guidelines, best practices and information security standards. There is a rich literature describing ontologies to represent such rules for the security and privacy management area. Abou-Tair [10] presents a way to enforce privacy in enterprises using ontologies to generate XACML [11] policies. The work presents the BDSG (Federal Legislation on Data Protection) ontology in F-Logic mapping law statements to a machine interpretable language. In our work the integration of resources to support the maintenance of a KB on the privacy domain establishes a space for common understanding necessary for the implementation of privacy rules in accordance with legal constraints and local policies among others.

Hecker [12] argues that privacy ontologies must show different concepts and associations, enabling interoperability and determining the privacy level of a transaction. Ontologies can also guide system developers who need to implement privacy functionalities or mechanisms without requiring expertise from developers specialized in the privacy domain. The proposed integration of resources on the basis of an ontology aims at the integration between system developers and other stakeholders in the requirements elicitation task.

Hu [13] proposes that the semantic model for EPAL privacy policies [14] can be expressed as a variety of ontologies and rule combinations. It supports the idea that ontologies are the main body of concepts to establish an infrastructure for the knowledge management in a domain. Our work does not focus on rules. Instead, ontology concepts are mapped to resources in the domain to support the challenge of semantic representation. It defines the basis for the enforcement of privacy, as well as for knowledge management in the privacy domain.

Although there are many ontologies in the privacy domain, reusing them is a difficult task, as they are developed for a wide variety of purposes, which differ from the specifics of our context. Our privacy ontology was manually built, based on the study of regulatory documents, guidelines, and also on some aspects of the KB system. Ontology concepts are then a central structure, from which other knowledge and linguistic resources are generated (Noun Phrase taxonomy, thesaurus, and NE). Several domain documents are mapped to the enriched ontology. Identified concepts and their extensions are then linked to all the domain resources in which they occur. Therefore the privacy ontology serves as a guide to several knowledge related tasks in which domain stakeholders are involved.

3. Knowledge resources

A manually built privacy ontology, validated by a privacy officer, a lawyer, and a project manager, was enriched with other resources on the basis of corpus processing. These resources are composed of a thesaurus, a noun phrase taxonomy, and NE. The corpus-based thesaurus relates terms that are similar to each ontology concept, and constitutes an extra semantic resource for assisting stakeholders. NE guide the access to important law documents. The taxonomy shows concepts related noun phrases organized in a hierarchy, which helps gathering information about contextualization of terms.

The remaining resources support the inference of risks by the accountability tool, comprising a glossary that describes important terminology, a questionnaire and a rule set that guides the flow of questions for privacy assessment of projects, resulting in the inference of the project risk level. Relating the system KB to the enriched ontology and the corpus is considered as an aid for engineers responsible for system updates.

The domain knowledge resources can be accessed on the basis of a given term, selected in a visualization tool available at http://www.cpca.pucrs.br/VisualizationTool/. The Ontology is viewed as a hyperbolic tree of concepts, instances and related properties. Such a view of fundamental domain concepts is then integrated with all the other knowledge resources. Users can then access related concepts in the thesaurus, and from it, navigating through all the accountability tool resources. A resource can be accessed through its tab or through context menus. The following section explains in more detail each accountability tool resource.

3.1 Accountability Tool Resources

This section describes resources directly related to the accountability tool. To accomplish with project restrictions, these resources are presented as figures on the text, and omitted from the visualization tool available.

3.1.1 Glossary

The domain glossary can help clarifying terms to stakeholders, represented as an entry to which a description is given, or as part of the description. An occurrence of *personal information* can be shown in Figure 1.

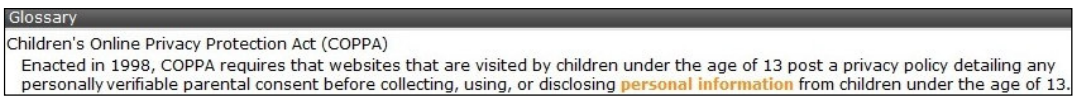


Figure 1 – Excerpt of the Glossary, with "personal information" as part of description

3.1.2 Inference Rules

Inference rules are managed by a risk inference component in the accountability tool. They guide the flow of questions that are shown, as well as determine the project privacy risks based on the answers provided by the user. They are structured as follows: rule name, risk indicator, origin of the rule, reason for the rule, remediation, and condition to fire the rule. As seen in Figure 2, when modeling requirements, KB engineers may learn that a notice statement must be provided by the system before collecting personal information, and also that other issues are involved. Similarly, privacy officers can check which rules will be affected when a change in the body of laws involving personal information occurs. Project managers can access occurrences of the term in the rulebook to help mitigate privacy risks for their projects as well as to manage organizational resources affected by rules related to personal information.

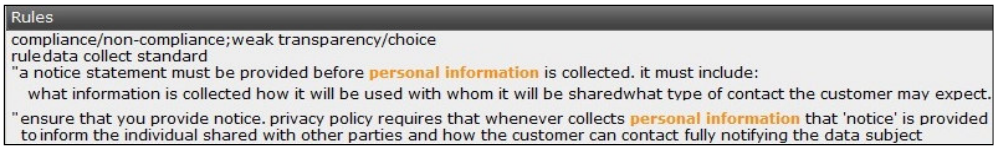


Figure 2 – Excerpt of the rules with the personal information concept

3.1.3 Questionnaire

The questionnaire is managed by the risk inference tool through questionnaire and compliance rules. The former rules involve pairs of questions and their possible answers, and also allow for more user-friendly grouping and ordering of questions. They are also used to set the value of intermediate variables, to decide which questions should be shown to the user, given the answers already provided. Intermediate variables are kinds of flags with a semantic meaning, hand-created by privacy experts to simplify the authoring of the rule base and manage the relation between the data comprised in questionnaire sections, and the knowledge it represents. Based on given answers a set of compliance rules with the form "*when condition then action*" infers the privacy compliance level of the project [15]. A compliance report is generated, with the results of the assessment of privacy risks and a list of remediations in case of higher-risk privacy concerns.

Figure 3 presents the term *personal information* in the question 66. For each question or answer with at least one occurrence of the mapped term the system presents the text of question and answer to better contextualize it. KB engineers can thus gain immediate and comprehensive control of impacts of domain changes to the questionnaire, and along with privacy officers keep the various objects in the rule base aligned with regulations and internally consistent during KB management.

Figure 3 – Excerpt of questionnaire with the personal information concept

The next sections describe the ontology and the remaining corpus-based resources.

3.2 Privacy Ontology

Despite the maturity in this field [16] reuse is difficult since each proposal is created for different purposes. Privacy risk assessment and analysis vary according to the

requirements imposed by specific scenarios. The definition of our Privacy ontology involves modeling concepts from several knowledge sources related to the problem of data privacy accountability, such as a set of legal documents, an accountability tool, and particular rules considered in the privacy risks inference scenario. The overall goal is the reduction of the difficulty of KB maintainability. To better understand the following explanation of the main concepts, we suggest the exploitation of the ontology through the visualization tool. Ontology concepts are represented as seeds on the Thesaurus tab.

Some concepts were chosen to identify references to legal documents. Thus, regulations are classified as normative and non-normative. Regarding the accountability tool, the ontology includes concepts related to project activities and purposes, user information, and sensitive information. Other essential concepts are *PII* and *Sensitive_PII*. People and organizations are also important concepts, because they refer to those involved in a transaction handling PII.

Concerning the idea of privacy risks, the ontology includes different risk levels. *Actions* conducted in a project can be associated to different *Risk Levels*. *Actions* and *activities* with no associated risks are evaluated to a *green level*. When internal policies are violated, the risk level associated with the activity is evaluated to a *yellow level*, and finally, the *red level* is attributed to activities that violate laws or regulations. Geographic locations, classified by the concept *Geo* as *cities*, *continents* and *countries* directly affect the definition of privacy risks.

All these concepts can be used in the description of project actions and their associated risks. In case of transborder data flows, for instance, risks depend on the kind of information, and on the origin and destination of the data flow.

The EU Data Protection Directive 95/46/EC, for example, imposes restrictions on the flow of PII to a third country, outside the European Economic Area [17]. A country is considered adequate for the flow of personal data if its laws provide a level of protection for personal data comparable to the Directive. Otherwise, it is considered non-adequate.

3.3 Corpus

The corpus used in our project was composed of a set of 100 documents of privacy regulations and development guidelines. By accessing the concept *personal information* in the visualization of the corpus, each occurrence of the term in a document is displayed in a context defined by five words on its left and right, which is called a concordance, along with the document identification, and the line number. This link can be used by Privacy Officers to evaluate how these concepts are used in regulations contained in the corpus, e.g., to verify that company practices (or the KB) are aligned with these regulations including in the presence of regulatory changes.

When a user selects the document name in the column Corpus, the original text file is highlighted in a concordance. A KB engineer can have a better understanding of requirements involving the flow of personal information to avoid the transfer of information without some adequate level of protection corresponding to Section 12, Item 1 of the highlighted text, for example. Similarly, a project manager can browse through the corpus of laws and guidelines to discover which documents can affect a system update involving, for example, additional transborder data flows of personal information. Stakeholders can inspect the KB, in order to decide the implications of changes to their respective fields in the organization.

KB engineers can more effectively model requirements involving rules and laws or check for KB correctness through being aided in the interaction with privacy officers by searches in the corpus. The inspection of other resources also helps to clarify a term in the domain, and to become aware of the impact of lawsuits arising from the misuse of personal information among others as seen in the following sections.

3.4 Corpus based Ontology Enriching Resources

The next subsections describe the resources extracted from the corpus directly related to domain concepts.

3.4.1 Thesaurus

As legal documents have large quantities of domain specific terms whose meaning can be represented with different terms the creation and maintenance of a thesaurus is a task that requires technological support. A thesaurus is composed of terms called seeds, to which similar terms in the domain are related. Associating a thesaurus to an ontology, and to a domain corpus can increase the efficiency of document retrieval. Instead of retrieving only documents containing specific terms the ones with terms semantically related can be retrieved. For example considering the term *personal_information* it is also referred to in the corpus as *personal_identifiable_information*, *personal_data*, and as the acronym *pii*. Thus by associating a thesaurus to our privacy ontology instead of retrieving only documents that contain the occurrence of some specific term documents containing also related terms can be the retrieved enriching the results with semantic privacy meaning.

Each ontology concept represents a seed in the thesaurus. To each seed shown on the tab "concept" of the visualization tool, a list of related terms from the corpus on the right was automatically generated using linguistic and statistical techniques. The ontology concept *personal_information* is found as similar to the terms, *PII*, *patient record* and *sensitive information*. By choosing a term in the thesaurus its occurrences in other knowledge resources can be accessed by stakeholders.

3.4.2 Named Entities (NE)

NE can be used to populate an ontology with instances extracted from the domain terms. The automatic recognition of NE from legal and normative documents can help the construction of a conceptual base of the privacy domain. In our work NE from legal texts representing instances of classes that contain as keywords the terms *act*, *law*, and *rule* were used to populate the ontology [18]. A list of classes extracted from the corpus of laws is shown in Table 1.

Table 1 - Examples of classes extracted with NER

Original classes	Derived Classes
Act	Enactment, Number, Turn, Routine, Deed, Bit
Law	Police, Jurisprudence, Constabulary
Rule	Ruler, Normal, Pattern, Prescript, Regulation, Principle, Convention, Formula, Dominion

In the visualization tool the term *personal information* can be viewed along with some recognized NE with the class to which they belong (Act), and the name of the legal instrument that contains each term.

When privacy officers and KB engineers are involved in clarifying legal implications that may affect the definition of requirements involving the protection of

personal information, for example, the identification of NE related to the selected term can provide a list of legal regulations to be investigated. Also the NE classes which are represented by ontology concepts can be identified helping the investigation of conceptual constraints in modeling decisions. Project managers can investigate relations between *personal information* and the laws relating to it, through references to legal documents retrieved on the basis of the term to evaluate possible legal implications of this term on project risks, for example.

3.4.3 Taxonomy

A hierarchy of noun phrases related to domain concepts may help stakeholders with a broader view of the context in which ontology concepts occur in documents. This involves more complex structures, which are not modeled as ontology concepts or instances. The taxonomy can help KB engineers to understand the uses of ontology concepts, thus providing extra information about the domain through the inspection of the contexts in which the term occurs. Our taxonomy was developed by parsing the corpus to extract noun phrase hierarchies. This extraction is based on the identification of noun phrases, and for each one on the identification of its constituents and nucleus. Noun phrases with the same nucleus were grouped and organized in a hyperbolic tree according to its constituents.

The resources described up to now summarize the mapping of textual sources, on the basis of a term for the information handled by domain stakeholders. The following section explores the evaluation performed by domain experts in the thesaurus and in the NE, to validate our efforts aimed towards establishing an infrastructure to the KM in this domain.

Apart from the taxonomy, the previous resources serve as the basis for the annotation of all the other resources for the support of the KM conducted by domain stakeholders as following described.

4. Integration of Heterogeneous Knowledge Resources

The resources mapping task consists in the XML-based indexing of terms from the enriched ontology that occur in a set of domain related documents. The whole process includes the generation of new resources and the annotation of documents in a mapping model shown in Figure 4.

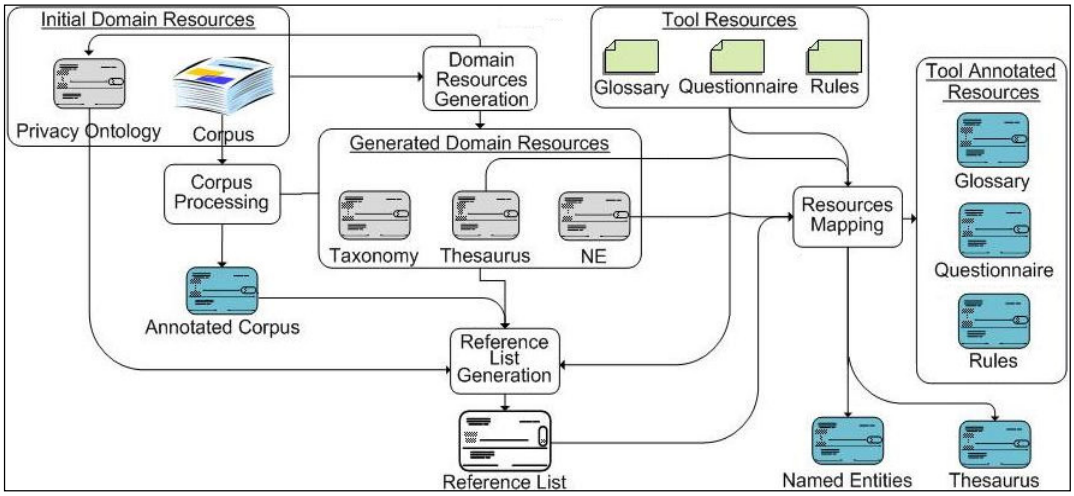


Figure 4 - Resources Generation and Mapping

The mapping process comprises the following steps:

- Generating thesaurus, taxonomy and NE from the ontology and the domain corpus.
- Generating a reference list of terms by merging the domain resources and the ontology concepts and instances.
- Checking the frequency of terms in each knowledge resource for the generation of the reference file.
- Annotating/indexing each knowledge resource based on the reference list.

The mapping procedures must always be performed when the domain KB is updated to set up new relations between resources. The indexing of the term *personal_information* can be seen in Figure 5 which shows the reference list of terms with the presence of the term in each resource.

```
<!-- reference.xml file>
<terms>
    . . .
    <term id="1124">
        <name><![CDATA[personal_data]]></name>
        <resource id="1" name="ontology" term_occurs="false" frequency="0"/>
        <resource id="2" name="thesaurus" term_occurs="true" frequency="1"/>
        <resource id="3" name="corpus" term_occurs="true" frequency="1349"/>
        <resource id="4" name="ne" term_occurs="false" frequency="0"/>
        <resource id="5" name="questionnaire" term_occurs="true" frequency="1"/>
        <resource id="6" name="glossary" term_occurs="true" frequency="1"/>
        <resource id="7" name="tagging" term_occurs="false" frequency="0"/>
        <resource id="8" name="rules" term_occurs="false" frequency="0"/>
```

Figure 5 – Excerpt of the reference XML file

All the other resources are represented by their specific XML files relating the presence of the term in it by its identifier, and specific attributes like the document number and line in which it occurs in the corpus.

4.1 Evaluation of the Ontology Enriching Methods

The evaluation process in our work consisted in verifying the quality of thesaurus generation and precision, recall and coreference for the NE recognition. Although tests for the evaluation of the overall integration of resources were not performed we performed the evaluation of thesaurus and NE, the most important resources that directly affect the enrichment of the domain ontology.

The evaluation of the thesaurus generation was performed by domain specialists including a privacy officer, a lawyer, and a project manager for a sample containing 10 domain concepts and 90 similar terms. The chosen concepts were: *children*, *consent*, *customer*, *data_protection*, *data_subject*, *marketing*, *notice*, *personal_data*, *personal_information*, and *regulation*. To evaluate them, specialists could assign a term as "*similar*", "*not similar*", or "*not sure*" (about the similarity). A term can also be ranked through arrows changing its position in the similarity list. A higher position on the list indicates a higher similarity level.

The precision rate for the sample of similar terms in the evaluation was 51.1%. We cannot fairly compare our results with related work because we do not share the same data. In practical terms, the production of a list of related terms in which about half is likely to be considered useful (as in the case of our methods over our corpus) is an important aid for the knowledge engineering processing.

For the NE recognition three classes of *Normative_Regulation* were considered, namely *Act*, *Law* and *Rule*. Other classes were generated from them, as follows: *Act* (*Enactment*, *Number*), *Law* (*Police*, *Constabulary*) and *Rule* (*Prescript*, *Regulation*,

Principle, and *Convention*). For instance, the class *Number* resulted from the NE *New Tax System (Australian Business Number) Act 1999.*

The Privacy corpus was tagged for these NE. The tagging task resulted in 4863 references to NE and 1191 unique entities in the domain. An evaluation tool analyses the tagging output against a manually tagged reference to obtain precision, recall and F-measure for:

a) Unique entities, represented by unique references to entities names;
b) Repeated references to the same entities.

The evaluation of the NE recognition performed on the corpus found 389 out of 1191 unique entities. An amount of 1460 references out of 4863 were found [18]. Resulting measures including precision, recall and F-measure are presented in Table 2. The results were considered promising and comparable to the results obtained from the 2008 ACE Local ERD. However, the application of more sophisticated natural language processing techniques over larger corpora can improve our results, in particular the recall measure [18].

Table 2 - NER processing resulting measures

	Precision	Recall	F-Measure
References to entities	60.48% (1460 / 2414)	30.02 % (1460 / 4863)	40.13
Unique entities	40.06 % (389 / 971)	32.66 % (389 / 1191)	35.99

We also evaluated coreference (or "*same_as*" relations), based on the search of acronyms. Table 3 has a row that represents both "*Employee Retirement Income Security Act of 1974*" and "*ERISA*", as they were found in the corpus as legal NE, and the system identified them as referring to the same entity. *ERISA* is said to be an acronym of "*Employee Retirement Income Security Act of 1974*". The evaluator was supposed to determine if this relation is correct or not, for the 185 instances related to it.

Table 3 – Acronyms for the relation "*same_as*"

Class	NE	Relation	Class	NE
Act	Employee Retirement Income Security Act of 1974	same_as	Act	ERISA
Act	TCPA	same_as	Act	Town and Country Planning Act 1990
Act	TCPA	same_as	Act	Telephone Consumer Protection Act of 1991

The evaluation of the relation "*same_as*" is presented in Table 4 according to 2 evaluators. We believe that the extraction of semantic relations between the entities recognized in this work and those which relate region-specific laws to their specific geo-political units can improve these results [18].

Table 4 - Evaluator's results for the relation "*same_as*"

	Evaluator 1	Evaluator 2
Correct	52.97 % (98 / 185)	67.03 % (124 / 185)
Incorrect	47.03 % (86 / 185)	32.97 % (60 / 185)

Concerning the taxonomy an evaluation was not performed since the resulting structure is just a straightforward reorganization of syntactic structures. However, the taxonomy generation tool [19] was previously evaluated in [20].

The enriching techniques developed so far can be considered as semi-automatic processes, whose output must be checked by experts given that the error rates are still considerably high. Suggestions of terms are provided by these techniques but an expert is needed in order to approve or refuse these suggestions. However these areas of NLP

are still under development and it is likely that the near future will bring new techniques with better recall and precision.

5. Concluding Remarks

Our work describes an ontology-based integration of knowledge resources in the privacy domain to support an accountability tool, focusing on the definition of concepts and the automatic enrichment of a privacy ontology, and on the construction and mapping of knowledge resources to support KM in the domain. The generation of relations between ontology concepts and various knowledge sources established the basis for knowledge inspection and refinement in accordance with changes in laws or in policies and requirements of the organization. The impact of such changes on the resources can be evaluated with the help of the integrated visualization tool developed in the project.

The domain concepts defined in the privacy ontology can be used to support the maintenance of the accountability tool. Our efforts were aimed at the definition of the mapping structure to integrate domain resources, and at the deployment of a tool to permit stakeholders to explore the knowledge and evaluate impacts of changes in the domain. As a result our ontology is composed of 113 concepts and 268 instances.

These efforts resulted in a semantic support that can help navigate through several resources and documents. The generated thesaurus can help specialists to identify similar terms for information search. NE are useful to keep track of changes in laws that need to be considered for KB maintenance. The integrated visualization of knowledge resources can help finding terms in a vast corpus of laws and other domain documents on the basis of an enriched ontology.

The Privacy ontology could not be fully reused for the management of privacy in different companies because it was defined to support a specific accountability tool, and it refers to concepts and instances of Project Activity and User modeled according to specific requirements. However, most concepts will remain useful in an ontology engineering process for similar ontologies.

We consider exploring more specialized semantic relations and features for the automatic recovery of information as future work.

References

1. Yee, G.O.M.; Korba, L.; Song, R. (2008) “Cooperative Visualization of Privacy Risks”, In: 5th International Conference in Cooperative Design, Visualization and Engineering, LNCS, vol. 5220, pp. 45-53.
2. Weitzner, D. J.; Abelson, H.; Berners-Lee, T.; Feigenbaum, J.; Hendler, J.; Sussman, G. J. (2008) “Information accountability”, *Commun. ACM* vol. 51, no. 6, pp. 82-87.
3. Mont, M.; Thyne, R. (2006) “Privacy policy enforcement in enterprises with identity management solutions”, In: PST '06, vol. 380, pp. 1-12.
4. Solove, D. J. (2006) “A Taxonomy of Privacy”, *University of Pennsylvania Law Review*, vol.154, no. 3, p. 477.
5. Rachamadugu, V.; Anderson, J. A. (2008) “Managing Security and Privacy Integration across Enterprise Business Process and Infrastructure”, In: 2008 IEEE Intl. Conf. Services Computing, vol. 2, pp. 351-358.

6. Knutson, T. R. (2007) “Building Privacy into Software Products and Services”, *IEEE Security and Privacy*, vol. 5, no. 3, pp. 72-74.

7. Duncan, G. (2007) “Engineering: Privacy by Design”, *Science*, vol. 317, n. 5842, pp. 1178-1179.

8. Ye, X.; Zhu, Z.; Peng, Y.; Xie, F. (2009) “Privacy Aware Engineering: A Case Study”, *Journal of Software*, vol. 4, no. 3, pp. 218-225.

9. Schäfer, U. (2006) “OntoNERdIE-Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources”, Proceedings of the 5th International Conference on Language Resources and Evaluation LREC - Volume 07.

10. Abou-Tair, D.D.; Berlik, S.; Kelter, U. (2007) “Enforcing Privacy by Means of an Ontology Driven XACML Framework”, In: IAS 2007, Third International Symposium on Information Assurance and Security, pp. 279-284.

11. OASIS XACML Technical Committee. (2003) “eXtensible Access Control Markup Language”.

12. Hecker, M.; Dillon, T. S.; Chang, E. (2008) “Privacy Ontology Support for E-Commerce”, *IEEE Internet Computing*, vol. 12, no. 2, pp. 54-61.

13. Hu, Y.; Guo, H.; Lin, A. G. (2008) “Semantic Enforcement of Privacy Protection Policies via the Combination of Ontologies and Rules”. In: IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pp. 400-407.

14. Backes, M.; Pfitzmann, B.; Schunter, M. (2003) “A toolkit for managing enterprise privacy policies”, In: ESORICS’03, vol. 2808, pp. 162-180.

15. Pearson, S.; Rao, P.; Sander, T.; Parry, A.; Paull, A.; Patruni, S.; Dandamudi-Ratnakar, V.; Sharma, P. (2009) “Scalable, Accountable Privacy Management for Large Organizations”. In EDOCW 2009, pp. 168-175.

16. Cybenko, G. “Why Johnny can't evaluate Security Risk”. *IEEE Security and Privacy*, vol. 4, no. 1 , p. 5.

17. European Commission: Commission decisions on the adequacy of the protection of personal data in third countries. From the Internet, accessed in 11/25/2010 on the URL: http://ec.europa.eu/justice/policies/privacy/thridcountries/index_en.htm

18. Bruckschen, M.; Northfleet, C.; Silva, D. M.; Bridi, P.; Granada, R.; Vieira, R.; Rao, P.; Sander, T. (2010) “Named Entity recognition in the legal domain for ontology population”, In: LREC 2010, pp. 16-21.

19. Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. (2009) “ExATOlp - An automatic tool for term extraction from portuguese language corpora”, In: LTC’09 - 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics.

20. Lopes, L.; Oliveira, L.; Vieira, R. (2010) “Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches”. In: PROPOR 2010 - International Conference on Computational Processing of Portuguese Language.

Part II
Short Papers

Proposta de uma Arquitetura para o Gerenciamento de Regras de Negócio em LPS com Base na MDA

Jaguaraci Batista Silva

Laboratório de Sistemas Distribuídos – Universidade Federal da Bahia – Salvador, BA

jaguarac@ufba.br

Abstract. *Are upheld in the SPL domain all business rules that define semantics links between entities for its completeness. By any change in business needs implies to keep of these rules up-to-date and is necessary to fix several application code to realize manually the necessary repair, making it a hard work for developers. The purpose of this works to create a new boarding to support software products in front of constant changes of business rules in this scenario, with thought at the reference architecture to manage them.*

Resumo. *Fazem parte do domínio de um uma LPS todas as regras de negócio que definem os laços semânticos entre entidades para a sua completude. Qualquer mudança nos requisitos de negócio implica na manutenção dessas regras e é preciso percorrer todo o código para realizar as mudanças necessárias manualmente, tornando esse trabalho árduo para os desenvolvedores. A proposta deste trabalho é criar uma nova forma de manter produtos de software frente às trocas constantes de regras de negócio neste cenário, com o suporte de uma arquitetura para gerenciá-las.*

1. Introdução

Há tempos, pessoas envolvidas com a fabricação de software vêm criando consciência de que para a satisfação dos clientes, é necessário construir produtos com qualidade. Porém não basta ter uma boa qualidade, se não for economicamente viável [Travassos *et al*, 2002]. A engenharia de domínio, bem como, outras técnicas visa à reutilização e está entre as técnicas mais relevantes para que se possa construir um software em menor tempo, maior confiabilidade e tendo como conseqüência um menor custo.

A Arquitetura Orientada a Modelos ou MDA (*Model-Driven Architecture*) [OMG, 2009] é uma abordagem de desenvolvimento de software dirigido por modelos (MDD ou *Model-Driven Development*) [Bragança e Machado, 2007] que utiliza modelos em diferentes níveis de abstração visando separar a arquitetura conceitual do sistema do seu modelo específico de plataforma. A definição de mecanismos de transformação permite a tradução dos modelos em código de linguagem de alto nível que, por sua vez, são compilados para geração do código executável da aplicação, facilitando o reuso e a manutenção das aplicações. Tais vantagens têm propiciado a crescente adoção da MDA como *framework* de desenvolvimento em diversos domínios de aplicação nos últimos anos.

Apesar dos avanços da comunidade científica na área de construção e reuso de modelos MDA [OMG, 2009] as empresas que trabalham utilizando linhas de produtos de software precisam construir um domínio de aplicação contendo as mesmas regras de negócio em várias plataformas, muitas vezes para prover funcionalidades similares em plataformas específicas. Qualquer mudança nos requisitos de negócio implica em esforço e custo de manutenção dos modelos específicos e independentes de plataforma. Para garantir que as regras de negócio ou propriedades específicas do domínio sejam atendidas pelo modelo gerado é comum aos desenvolvedores inserir o código que representa cada regra diretamente no modelo específico de plataforma (PSM) [OMG, 2009], o que tende a ser fastiento e sujeito a erros. Assim, a falta de mecanismos facilitadores que permitam a garantia de atendimento a essas regras, frente às constantes trocas de requisitos de negócio, torna-se um empecilho para a construção de aplicações confiáveis e de fácil reuso e manutenção.

Este trabalho propõe uma arquitetura para o gerenciamento eficiente de regras de negócio a partir de uma abordagem top-down usando a arquitetura orientada a modelos (MDA). A seção 2 e 3 apresenta os trabalhos relacionados, as pesquisas realizadas até o momento e os seus resultados. A seção 4 aduzi a respeito da arquitetura proposta. Por fim, as conclusões, contribuições e referências estão na parte final deste artigo.

2. Trabalhos Relacionados

A aproximação da MDA com uma família de LPS é um tema relevante e emergente onde se busca reduzir os custos de produção de software [Braga et al, 2007]. Em 4SRS ou *Four Step Rule Set*, foi concebido um framework utilizado pela Universidade do Minho. A sua finalidade é integrar partes de diferentes trabalhos na área de MDA e LPS visando à obtenção de regras de negócio a partir dos diagramas de casos de uso e derivando-as para diagramas de classe da UML [Braga et al, 2007].

Da aplicação conjunta da MDA em uma LPS também pode se beneficiar do melhoramento da gerência das variabilidades do domínio, sendo possível uma solução para o problema de sincronia entre os modelos específicos e independentes de plataforma. Um dos possíveis resultados seria a evolução das propriedades do domínio: conceitos, relações e restrições dentro de uma família de LPS de forma independente da arquitetura de implementação [Deelstra et al, 2003]. Em um dos trabalhos pesquisados, foi criada uma ontologia de domínio para modelar variabilidades de um software de apoio à configuração que inclui: componentes e *features* com suporte as estruturas de composição e seus atributos, interfaces, conexões e restrições. A ontologia utilizou a linguagem natural e um perfil UML com o suporte às inferências sobre o domínio para fornecer verificação semântica formal dos componentes [Asikainen et al, 2006].

Outra questão relevante é a necessidade de se prover técnicas e ferramentas para que se possa ter agilidade na tradução da lógica de negócios. Alguns trabalhos permitem a integração com ferramentas existentes no mercado, o que facilita o uso dessas técnicas nas empresas. Joukhadar (2008) criou um framework baseado em perfis da linguagem UML e utiliza diagramas de classes para capturar os conceitos, suas propriedades e relações e provê a especificação de regras de negócios em linguagem natural, de forma automática sem a ajuda do programador. Em uma pesquisa recente foi criada a

plataforma *SMICE* [Wu et al, 2010] baseada na MDA com objetivo de envolver tanto o cliente como os desenvolvedores na construção de softwares para a WEB. A plataforma oferece uma ferramenta para desenho dos componentes básicos dos processos, suas integrações e regras de negócio. O cliente define todo processo utilizando qualquer software específico para BPM [Santos et al, 2011] e a ferramenta converte o processo e suas regras de negócio do formato BPMN para WS-BPEL, simplificando a criação e manutenção de serviços na Web.

3. Resultados

Durante a construção de uma linha de produtos é preciso também ter um processo conciso para construir e verificar os pontos de variabilidade do domínio e a falta de um conjunto detalhado de atividades e ferramentas para apoiar esse processo foi um dos principais problemas enfrentados na pesquisa. Em um trabalho anterior Silva *e Saba* (2008) propuseram a criação de um processo de integração de componentes com base no guia *CMMI [SEI, 2002]* e uma metodologia de *Business Process Management* (BPM), que mescla a padronização de documentos, diagramas e representação de atividades junto com técnicas de modelagem de requisitos baseadas em casos de usos. A técnica de extração de requisitos com base em casos de uso foi utilizada por ser bastante difundida nas empresas desenvolvedoras de software. Aproveitando essa padronização foi criada uma aproximação com ontologias para a concepção de um processo a fim de guiar a construção, verificação e validação de domínios de aplicação [Silva, 2010].

A transformação de modelos MDA com base em ontologias de domínios requereu aprofundar os estudos sobre quais ferramentas e técnicas poderiam ser utilizadas para permitir a sua construção e verificação semântica. Silva e Pezzin (2007) definiram um toolkit utilizando ferramentas do mercado combinando o uso de modelos nas linguagens OWL e UML. Esse trabalho possibilitou a transformação do modelo conceitual ou ontologia de domínio no formato OWL (PIM) em um modelo para banco de dados (PSM). Realizando uma verificação e validação semântica do domínio antes da transformação dos modelos através de raciocínio automático, usando lógica de descrição sobre o arquivo OWL gerado a partir da ontologia [Haarslev, 2004].

Durante as inspeções de código foi constatado que os axiomas ou regras de negócios provocavam um alto grau de acoplamento, dificultando a sincronia, evolução e manutenção dos modelos. Por serem convertidas normalmente para estruturas de controle de uma linguagem de programação (e.g. IF-Then-Else), tornavam-se muitas vezes repetitivas. A proliferação de muitos pontos de controle de regras de negócio, frente às constantes trocas de requisitos, tornou-se um trabalho árduo para os desenvolvedores. Uma solução para este problema foi criar um processo para separação e validação de regras de negócio [Silva e Barreto, 2008] e a ferramenta *OWLtoASPECTJ* para realizar esta tarefa de forma automática [Silva, 2008]. A ferramenta transforma os axiomas, após serem verificados e validados semanticamente, em um novo modelo específico de plataforma (PSM) contendo apenas as regras de negócio. Através da programação orientada a aspectos [Laddad, 2003], as regras encontram dinamicamente e em tempo de execução, os métodos a serem interceptados, tornando possível a verificação da mesma regra em vários pontos da aplicação. Além disso, as regras geradas na linguagem *AspectJ* podem ser reutilizadas em outras plataformas que

utilizam a linguagem Java [Soares *et al*, 2002] agregando valor ao uso da MDA em uma família de LPS.

4. Proposta para o Gerenciamento de Regras de Negócio MDA em LPS

A Figura 1 apresenta a proposta em forma de diagrama de componentes da UML para facilitar a compreensão. A arquitetura ou ambiente de construção e verificação de modelos MDA será composto por 4 componentes: 1 - Ferramenta de Construção de Domínio, 2 - Ferramenta de Verificação Semântica de Regras de Negócio, 3 - Ferramenta de Separação do Domínio da aplicação e 4 - Repositório de Regras de Negócio ou Base de Conhecimento.

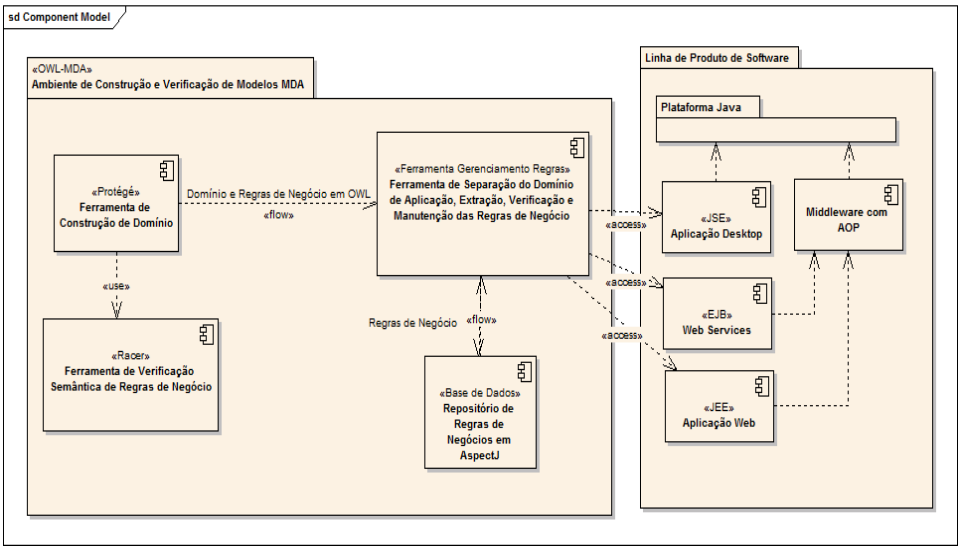


Figura 1 – Arquitetura Proposta para o Gerenciamento das Regras de Negócio em LPS usando MDA.

A ferramenta de edição de ontologias representada pelo componente <Protégé> [Protégé, 2011] será designada para a construção da ontologia de domínios da LPS. O primeiro passo é a criação do domínio conceitual da aplicação contendo: as entidades, suas propriedades e relacionamentos e os axiomas. O artefato base produzido por esse componente será uma ontologia de domínio consonante com os requisitos da aplicação. A construção da ontologia será guiada a partir de casos de uso que identificam os requisitos funcionais de uma família de LPS [Silva e Barreto, 2008].

O domínio conceitual da aplicação ou ontologia de domínio deverá sofrer inferências para a sua verificação e validação através da ferramenta de verificação semântica de regras de negócio, representado na Figura 1 pelo componente <Racer> [Haarslev, 2004]. Após serem validados, os axiomas (regras de negócios) serão transformados do formato OWL (*Ontology Web Language*) (PIM) para uma linguagem de programação orientada a aspectos (PSM) com a ajuda da ferramenta de separação do domínio da aplicação ou componente <Ferramenta de Gerenciamento de Regras> OWLtoAspectJ [Silva, 2008].

As regras de negócio extraídas da ontologia de domínio farão parte de um único modelo (PSM) que servirá para o tratamento da lógica de negócio e serão armazenadas em uma base de conhecimento ou repositório de regras de negócios, reproduzido na gravura pelo componente <Base de Dados>. Os outros modelos serão criados de acordo

com as necessidades não-funcionais dos casos de uso. No exemplo da LPS proposta na Figura 1, seria necessário gerar 3 modelos específicos para a plataforma Java: o primeiro para compor uma aplicação que irá executar em computadores desktops <JSE>, o segundo para uma aplicação Web <JEE> e o último serviços Web <EJB>.

A arquitetura deverá provê o suporte à atualização das regras de negócios de forma centralizada e automática. A ferramenta de separação do domínio da aplicação fará a extração das novas regras da ontologia corrigida (Seção 3), realizará o merge na base de conhecimento das regras antigas e atuais, realizando busca semântica através de taxonomias e finalizará com a inserção das novas regras na LPS.

Com a utilização do paradigma de programação orientada a aspectos [Kiczales, 1996], os modelos de regras e específicos de plataforma podem ser combinados para que a aplicação possa interceptar automaticamente as regras em tempo de execução sem que o desenvolvedor necessite conhecer os pontos de reparação do software (Seção 3). Seja usando diretamente a linguagem AspectJ, componente <JSE>, ou um *Midleware* com o seu suporte, componentes <EJB> e <JEE>.

5. Conclusão

Gerenciar as regras de negócio para um produto de software não é trivial e tal dificuldade aumenta substancialmente em uma linha de produtos de software. Neste âmbito, pensar nos modelos MDA para apoiar a gerência de tais regras torna-se uma alternativa relevante. O artigo propõe uma arquitetura para melhorar o gerenciamento de regras de negócio no contexto de linhas de produtos de software, utilizando uma abordagem MDA. Foram definidos os componentes que integram essa arquitetura, bem como, os resultados alcançados até o momento. Espera-se em uma próxima etapa validar a arquitetura em um estudo de caso real de um sistema para uma grande instituição do mercado financeiro usando as tecnologias Java (e.g. JSE, EJB e JEE) [Sun, 2011] por já ser bastante utilizada no mercado.

Referências

Asikainen,T.,Mannisto, T.,Soininen, T.. (2006) “Kumbang: A domain ontology for modelling variability in software product families”. Advanced Engineering Informatics, Vol. 21 (2006) 23–40, Springer.

Bragança, A. Machado, R. J. (2007) “Model Driven Development of Software Product Lines”. IEEE 6th International Conference on The Quality of Information and Communication Technology (QUATIC), 2007, Lisboa, Portugal.

Desltra, S. Sinnema, M., Gurp, J. V., Bosch, J.. (2003) “Model Driven Architecture as Approach to Manage Variability in Software Product Families”. Workshop on Model Driven Architecture: Foundations and applications, 2003, Enschede, Holanda.

Gimenes, I. M., Travassos, G. H.. (2002) O enfoque de Linha de Produto para Desenvolvimento de Software. In: Sociedade Brasileira de Computação; Ingrid Jansch Porto. (Org.). XXI JAI - Livro Texto. Florianopolis: Sociedade Brasileira de Computação, v. 2, p. 1-32.

Haarslev, V., Muller, R.. (2004) “Racer’s User Guide and Reference Manual”. Versão 1.7.19.

Joukhadar, A., Al-Maghout, H. (2008) "Improving agility in business applications using ontology based multilingual understanding of natural business rules", International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA 2008). Damascus, Siria.

Kiczales, G. (1996). "Aspect-oriented programming". ACM Computing Surveys, 28A(4), 1996.

OMG (2009). MDA Guide version 1.0.1. Formal Document: 03-06-01. Disponível em: http://www.omg.org/cgi-bin/apps/doc?omg/03-06-01.pdf. Acesso: Maio/2011.

Protégé. "Ontology Editor and knowledge-base framework". http://protege.stanford.edu/. Acesso: Maio/2011.

Laddad, R.. "AspectJ in Action, Practical Aspect-Oriented Programming". Manning,

ISBN 1-930110-93-6. 2003.

Santos, A.G., Santos F. G., Mendes F. A. T., Cruz G. M., Silva J. B., Freitas J. V. V. B., Santana M. R., Pastor S. O. (2006) "Metodologia de Processos de Negócios", Universidade Federal da Bahia, Programa de Residência em Software com foco em e-government, disponível em http://twiki.im.ufba.br/bin/view/Residencia/Trabalhos. Acesso: Maio/2011.

SEI CMMI (2002) "The Capability Maturity Model for Software". Version 1.1-CMU/SEI-2002-TR-012, March.

Silva, J. B., Saba, H. (2008). "Modelagem das áreas de Processo do CMMI usando uma metodologia de BPM e notações do SPEM". 34º Congresso Infobrasil TI & Telecom 2008, Fortaleza.

Silva, J. B., Barreto, L. P.. (2008). "Separação e Validação de Regras de Negócio MDA Através de Ontologias e Orientação a Aspectos". Simpósio Brasileiro de Componentes, Arquitetura e Reuso de Software 2008, Porto Alegre.

Silva, J. B.. (2010). "PROCEDA – Um Processo para Construção e Verificação Semântica de Domínios em uma Linha de Produtos de Software". Universidade Federal da Bahia. Relatório de Pesquisa, Agosto/2010.

Silva, J. B., Pezzin J. "The Formal Verification of an Application Conceptual Model Using MDA and OWL". World Congress on Engineering and Computer Science (WCECS 2007), San Francisco, 2007.

Silva, J. B. "OWLtoAspectJ: A Tool for Transformation from Conceptual Rules of Domain to Aspects". I Seminário de Pesquisa em Ontologia no Brasil, Niterói, 2008.

Soares, S. Borba, P. "Programação Orientada a Aspectos em Java". VI Simpósio Brasileiro de Linguagens de Programação, Rio de Janeiro, 2002.

Sun (2011). Sun Microsystems (2011). "Java Technology". http://www.java.sun.com. Acesso: Maio/2011.

Wu, M., Jin, C., Ying, J.. (2010). "SMICE: A Platform Supports Business Process Modeling and Integration". 2nd IEEE International Conference on Information Management and Engineering (ICIME2010). Chengdu, China.

Abordagens Estocásticas para Raciocinadores aplicáveis em Web Semântica

Juliano T. Brignoli[1], Denilson Sell[2], Fernando O. Gauthier[3]

[1]IFC - Instituto Federal Catarinense
Rio do Sul - SC – Brasil

[2]UFSC – Universidade Federal de Santa Catarina
Pós-graduação em Engenharia e Gestão do Conhecimento
Florianópolis – SC – Brasil

[3]UFSC – Universidade Federal de Santa Catarina
Pós-graduação em Engenharia e Gestão do Conhecimento
Florianópolis – SC - Brasil

juliano.brignoli@gmail.com, denilson@stela.org.br, gauthier@inf.ufsc.br

Abstract. *We propose a treatment of uncertainty in areas of Semantic Web ontologies for modifying the schema PR-OWL.O scheme is limited to considering the uncertainties of randomness, but certain areas of representation of the real world can also have imprecise variables. To qualify for the process examined plausible inference is a hybrid model of reasoning that can theoretically be incorporated into the scheme of syntactic and functional PR-OWL. The article explores this conceptual integration, assuming that PR-OWL will increase the accuracy of the results of inference.*

Resumo. *Propõe-se um tratamento de incertezas em domínios de ontologias para Web Semântica modificando-se o esquema PR-OWL. O esquema limita-se a considerar as incertezas por aleatoriedade, mas, certos domínios de representação do mundo real também podem apresentar variáveis imprecisas. Para qualificar o processo de inferência plausível analisou-se um modelo de raciocínio híbrido que, teoricamente, pode ser incorporado no esquema sintático e funcional da PR-OWL. O artigo explora conceitualmente esta integração, pressupondo que PR-OWL aumente a acurácia dos resultados da inferência.*

1. Introdução

A inferência é um processo inerente à operacionalização das ontologias em Web Semântica, amplamente explorado em aplicações focadas na determinação de respostas a partir de uma busca analítica sobre um ambiente com variáveis interrelacionadas.

A teorização acerca da utilização de inferências perpassa fundamentalmente pelos postulados da Lógica de Primeira Ordem e todo o seu desenvolvimento aprimorado por séculos, dadas as contribuições de matemáticos e filósofos da antiguidade e da contemporaneidade.

No desenvolvimento da Web Semântica e, mais especificamente, nas construções de ontologias, os processos de inferência são executados por motores de raciocínio (*reasoners*). Existem diversas especificações ou formas de atuação dos raciocinadores aplicáveis em Web Semântica, existindo desde aqueles formulados por princípios da Lógica, por processos otimizados de busca com conotação semântica e ainda, aqueles com funcionalidades inerentes ao tipo de tratamento estocástico de informações.

Este artigo objetiva apresentar uma discussão acerca da utilização do raciocínio plausível em Web Semântica, destacando a Probabilistic-OWL como um dos formalismos para aplicações em domínios de ontologias com incertezas. Pretende abordar de maneira teórica a possibilidade da contribuição de um modelo híbrido de raciocínio, combinando o processamento da imprecisão e da aleatoriedade.

2. Os raciocinadores para Web Semântica

Raciocinadores computacionais realizam operações lógicas aplicadas sobre bases de conhecimentos compostas por fatos e são amplamente utilizados nas ontologias, agindo de modo a executar regras que buscam alcançar respostas a um processo de inferência.

Ontologias em Web Semântica podem apresentar domínios complexos e dinâmicos e dispõem de uma linguagem formal que descreve representações, é a OWL, normatizada pelo W3C. No âmbito da OWL encontra-se a OWL DL, como sendo uma derivação que suporta inferência a partir da Lógica Descritiva. Horrocks (2002) define a Lógica Descritiva como uma denominação geral para uma família de formalismos de representação do conhecimento. Ao referenciar ferramentas para representação de ontologias é evidente o uso intensivo do software Protégé (PROTÉGÉ, 2010). Para este framework existem vários plug-ins disponibilizados, tais como, o raciocinador Pellet, da Clark & Parsia LLC, sendo um mecanismo de inferência especificado em OWL DL, desenvolvido em Java e de código aberto. Conforme observado em (PELLET, 2010), Pellet é uma implementação baseada no algoritmo de Tableaux (tablô), um método formal oriundo da Lógica de Primeira Ordem. O Pellet dá suporte à especificação de tipos de dados que podem ser definidos pelo usuário, incorpora heurísticas para detecção de ontologias OWL Full na tentativa de expressá-las com OWL DL e identifica axiomas que causam inconsistências entre conceitos.

Em (SILVA, 2008) é possível conhecer diversos mecanismos de inferência para servirem às ontologias e entre vários observou-se o Fuzzy DL. Este sugere uma abordagem mais exploratória e uma reflexão acerca da utilização de métodos focados ao processamento de incertezas existentes em domínios nas ontologias para a Web Semântica. Na seqüência deste artigo, propõe-se discorrer sobre o contexto estocástico que pode ser manifestado na representação de ontologias.

3. Fuzzy e a incerteza pela imprecisão

Em conformidade com a abordagem de (ROSS, 1995), a Lógica Fuzzy é um formalismo matemático para modelar problemas que apresentam variáveis munidas de incertezas inerentes

à imprecisão de seus valores. Na Lógica Clássica (bivalente), dado um conjunto A, um elemento x $\in$ A ou x $\notin$ A, em Fuzzy, um elemento possui uma medida de possibilidade num intervalo [0, 1]. A modelagem fuzzy sugere o uso de variáveis lingüísticas para descrever o domínio em representação. Estas variáveis estão associadas com objetos ou conceitos do mundo real possíveis de uma especificação quantitativa, tais como, velocidade, altura, temperatura, etc. Estas variáveis são segmentadas em intervalos denominados Conjuntos Difusos que podem, em circunstâncias, caracterizar certas sobreposições de conceitos, como, pouco, pouquíssimo, muito, muitíssimo, denotando imprecisão de conceitos na representação de domínios.

4. Raciocínio Plausível Bayesiano

O raciocínio bayesiano apresenta solução de inferência na modelagem de domínios cujas variáveis estão sujeitas ao acaso, onde o fator randomicidade é o tipo de incerteza manifestada. É propício enfatizar uma diferença fundamental entre a proposta da Lógica Fuzzy e do Raciocínio Bayesiano: em Fuzzy, uma variável é incerta pela sua imprecisão e a certeza é uma medida de possibilidade. No contexto da aplicação do Raciocínio Bayesiano, as variáveis são incertas devido à aleatoriedade intrínseca e a certeza é uma medida de probabilidade.

Pearl (1988) aborda a descrição formal deste processo de inferência bayesiana onde a relação entre variáveis acarreta em uma rede de interdependências, logo, a modificação de uma variável do ambiente pode afetar várias outras em termos de crenças probabilísticas. De maneira sucinta apresenta-se o princípio da formulação matemática do Teorema de Bayes: Seja o espaço de probabilidade (ε, P) e os eventos compostos H1, H2,...,Hk $\subseteq \varepsilon$ desde que nenhum desses eventos tenha probabilidade nula, então:

$$P(Hi / e) = \frac{P(e / Hi) \cdot P(Hi)}{P(e)} \quad (1)$$

onde $P(e) = P(H1).P(e / H1) + P(H2).P(e / H2) + ... + P(Hk).P(e / Hk)$ (2)

As fórmulas (1) e (2) compõem o cerne do mecanismo de inferência das denominadas Redes Bayesianas. Contudo, conforme discorre Carvalho *et al.* (2008) estas redes apresentam limitações para representar diversas situações do mundo real e da Web como, dificuldades em lidar com elevado número de variáveis aleatórias e o uso de recursão, o que retrata problemas no campo da complexidade computacional. Logo, Carvalho *et al.* (2008) apresenta uma proposta acerca do raciocínio bayesiano em ontologias para Web Semântica utilizando Redes Bayesianas Multi-Entidades (MEBN). Estas redes integram a Lógica de Primeira Ordem com o raciocínio bayesiano, conforme proposta desenvolvida por Laskey (2008). No âmbito da especificação de ontologias, a linguagem OWL não suporta raciocínio plausível, assim, Carvalho *et al.* (2008) recorre à linguagem PR-OWL (probabilístic OWL) como sendo a primeira implementação de MEBN para ontologias de domínios com incertezas. (PR-OWL, 2011). Trata-se de uma contribuição para os trabalhos desenvolvidos pela URW3-XG (W3C Uncertainty Reasoning for the World Wide Web Incubator Group). As MEBN modelam conhecimento por meio de MFrags que constituem esquemas de entidades de domínio por meio de grafos. Segundo Carvalho *et al.* (2008), uma MFrag é uma distribuição de probabilidade das instâncias de suas variáveis aleatórias. A figura 1 ilustra por meio de uma MFrag, a modelagem que representa o nível de perigo a que uma nave é exposta.

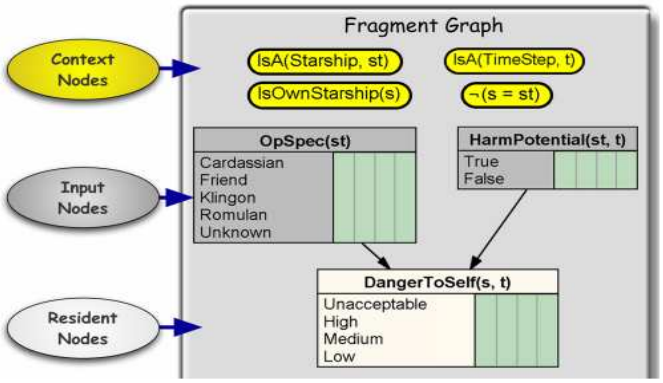


Figura 1 - exemplo de uma Mfrag (Carvalho *et al.* (2008))

O problema representado pela figura 1 é significativo para elucidar elementos de investigação no campo do processamento de incertezas. Observa-se a especificação de variáveis aleatórias que denotam subjetividade, ou seja, são vinculados a um domínio com valores imprecisos. Os estados *High, Medium, Low*, caracterizam conceitos difusos que incorporam mais incertezas à MEBN.

Em conformidade com esta observação objetiva-se convergir para uma proposição: utilizar Lógica Fuzzy na formulação híbrida do raciocínio plausível em ontologias para qualificar a inferência em função dos conceitos difusos existentes em nós residentes em uma MFrag. Na seqüência será apresentada a proposta de formulação.

5. Possibilidade da representação de Raciocínio Híbrido para Web Semântica

Em suas obras, Ross (1995) e Kandel (1986) introduzem algumas abordagens e ensaios matemáticos de modo a nortear pesquisas que venham a caracterizar uma utilização combinada dos processos de inferência fuzzy e bayesiana.

Brignoli (2001) sugeriu em seu trabalho que uma Rede Bayesiana pode ter uma redistribuição de suas probabilidades de saída (resultados) quando inferirem sob variáveis de entrada que apresentam imprecisão como incerteza eminente. Para tal, o autor apresenta $\phi = f(\rho, \delta)$ como uma função que expressa o que denominou de qualificador para o modelo híbrido de raciocínio. Mostra por simulação que ϕ causará a redistribuição das probabilidades da Rede Bayesiana tendo ρ como uma probabilidade condicional sem imprecisão e, δ representando um ou mais eventos difusos. Estes eventos podem ser as variáveis de entrada da Rede Bayesiana.

Brignoli (2001) utilizou a fórmula do Teorema de Bayes (1) com modificações em sua forma original, conforme segue:

$$P(Hi / \delta) = \frac{P(\delta / Hi)}{P(\delta)} \qquad (3)$$

onde,

$$P(\delta Hi) = P(Hi) \,.\, [\, P(\delta Hi) \,.\, \mu_{\mathcal{E}1}(\delta) + (1 - P(\delta Hi)) \,.\, \mu_{\mathcal{E}2}(\delta)] \qquad (4)$$

e,

$$P(\delta) = \sum_{i=1}^{n} P(Hi) \,.\, [\, P(\delta Hi) \,.\, \mu_{\mathcal{E}1}(\delta) + (1 - P(\delta Hi)) \,.\, \mu_{\mathcal{E}2}(\delta)] \qquad (5)$$

Na fórmula, $\varepsilon 1$ e $\varepsilon 2$ foram usados para representar conjuntos difusos de entrada que determinaram a incerteza por imprecisão existente nas variáveis de entrada da Rede Bayesiana. Em termos de confiabilidade do modelo é importante ressaltar que após ser aplicada a função ϕ de qualificação, o axioma estatístico $\sum P(Hi) = 1$ continua verdadeiro.

Uma conclusão notória obtida nesta modelagem é o que a utilização do modelo de inferência híbrido causou um espalhamento na distribuição de probabilidades da Rede.

> A argumentação para este espalhamento está baseada na observação de que, quando as probabilidades estão condicionadas a eventos difusos, aquelas acima do ponto de máxima entropia, diminuem, enquanto as demais aumentam. Desta forma, o processo de qualificação provoca uma redistribuição das probabilidades e o referencial para esta mudança de informação é justamente o ponto em que está situada a maior incerteza no conhecimento da Rede. (BRIGNOLI, 2001, p. 110)

Considerando a validade do modelo descrito, voltamos a referenciar a PR-OWL e o esquema ilustrado na figura-1. Carvalho *et al.* (2008) propôs a implementação de uma ferramenta visual que servisse como plug-in ao Protégé para apoiar a representação de ontologias com domínio probabilístico e incorporou funcionalidades da PR-OWL nesta ferramenta. O autor argumenta a vantagem desta implementação visual dado o fato que arquivos PR-OWL são similares aos arquivos HTML, sendo exaustiva sua construção e manutenção.

Conhecida e disponibilizada a PR-OWL, propõe-se a incorporação de mecanismos de inferência híbridos que agreguem aos raciocinadores de ontologias em Web Semântica, a partir, do próprio uso da PR-OWL, as funcionalidades para operar sobre domínios com imprecisão e aleatoriedade. Torna-se oportuno explorar uma esquematização para uma Hibrido-OWL como sendo uma extensão da PR-OWL. A figura 2 sugere um macro-esquema:

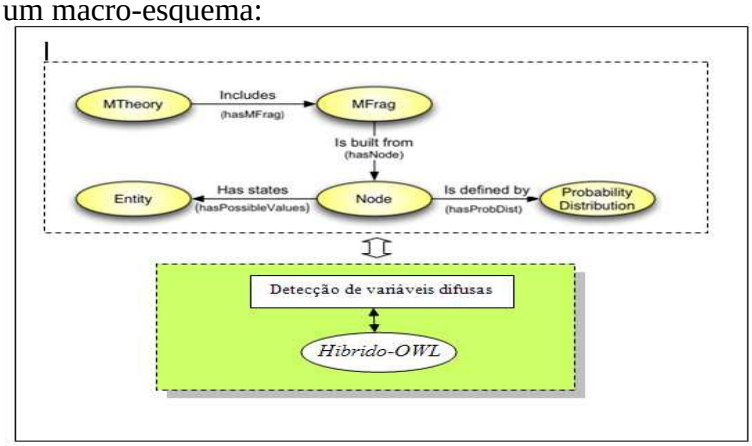


Figura 2 - esquema de uma PR-OWL híbrida (Adaptado de Carvalho et al. (2008))

É importante enfatizar que o objetivo é apresentar uma possibilidade de integração das funcionalidades de uma inferência híbrida fuzzy-bayesiana ao esquema da PR-OWL já em uso nas ontologias com domínios de incerteza.

Em termos operacionais e de implementação computacional, poder-se-ia agregar ao código PR-OWL a capacidade de inferir imprecisão em ontologias com a inserção adequada do modelo apresentado na fórmula (3).

6. Conclusões

A exploração teórica mostrou a possibilidade de uma investigação com maior profundidade no tema visando propostas implementáveis. A abordagem estocástica e o estudo das incertezas são convenientes no âmbito das ontologias em Web Semântica

devido à magnitude de variáveis que representam domínios. Notou-se que as ontologias no âmbito da Web Semântica podem acarretar conceitos e variáveis incertas, expressas na forma de imprecisão ou de plausibilidade.

O estudo das inferências foi fundamental para a proposta de construção de motores de raciocínio aplicáveis às ontologias em Web Semântica e a PR-OWL incrementou a qualidade da acurácia em ontologias que requerem, por sua natureza de descrição, uma forma de raciocínio plausível oriundo da aleatoriedade especificada nas relações das variáveis.

Na análise da funcionalidade da PR-OWL ficou perceptível sua limitação ao raciocínio sobre incertezas por imprecisão, ponto crucial que identificou uma lacuna no processo de raciocínio sobre incertezas nas ontologias em Web Semântica e oportunizou uma reflexão acerca da utilização de um raciocinador híbrido agregado às funcionalidades da PR-OWL. Por ter sido utilizado em outras aplicações, o raciocinador híbrido elucidado neste artigo sugeriu um caminho promissor na investigação sobre esquemas formais para lidar com cenários de incerteza em ontologias para Web Semântica. Esta verificação assegura, *a priori*, uma adequação ao esquema PR-OWL, com provável incremento da eficácia na inferência.

Como maior contribuição ao processo de inferência realizada em ontologias, o raciocinador híbrido argumentado neste trabalho tende em acarretar uma qualificação dos resultados no que tange ao aspecto da semântica produzida.

Referências

Bravo, Carlos de Oliveira. (2010). Geração Automática de Ontologias para Web Semântica. Dissertação de Mestrado. UnB. Brasília.

Brignoli, Juliano T. (2001). Modelo Híbrido Difuso-Probabilístico: uma alternativa para Sistemas Especialistas. Dissertação. UFSC. Florianópolis.

Carvalho, Rommel Novaes et al. (2008). Raciocínio Plausível na Web Semântica através de Redes Bayesianas Multi-Entidades – MEBN. Dissertação.UnB. Brasília.

Horrocks, I. (2002). DAML+OIL: a description logic for the semantic web. IEEE Data Enginee-ring Bulletin, 25(1):4–9. URL http://citeseer.ist.psu.edu/578691.html.

Kandel, Abraham. (1986). Fuzzy Mathematical Techniques with Applications. Florida: Addison-Wesley Publishing Company.

Laskey, K. B. (2008). MEBN: A Language for First-Order Bayesian Knowledge Bases. Artificial Intelligence, 172(2-3).

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. California: Morgan Kaufmann.

Pellet. (2010). Disponível em: http://pellet.owldl.org/. Acesso em:<15/08/2010>.

Protégé. (2010). Disponível em: http://protege.stanford.edu. Acesso em:<15/08/2010>.

Pr-OWL. (2011). Disponível em: http://www.pr-owl.org. Acesso em:<22/08/2011>.

Ross, Timothy J. (1995). Fuzzy Logic With Engineering Applications. McGraw-Hill.

Silva, Marcel Ferrante. (2008). Semantic web reasoning. Seminários de disciplina. Organização e Tratamento da Informação. UEMG. Belo Horizonte.

Hierarquias de Conceitos para um Ambiente Virtual de Ensino Extraídas de um *Corpus* de Jornais Populares

Maria José Bocorny Finatto[1], Lucelene Lopes[2], Renata Vieira[2], Aline Evers[3]

[1,3]Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul (UFRGS), [1]Pós-Doutoranda ICMC-USP
Av. Bento Gonçalves, 9500 – 91.540-000 – Porto Alegre – RS – Brasil

[2]Programa de Pós-Graduação em Ciência da Computação, Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Av. Ipiranga, 6681 – 90.619-900 – Porto Alegre – RS – Brasil

mfinatto@terra.com.br, lucelene.lopes@pucrs.br, renata.vieira@pucrs.br, aline.evers@gmail.com

Abstract. *In this paper we present conceptual hierarchies automatically obtained from the popular newspaper Diário Gaúcho. The hierarchies were generated by means of ExATOlp tool through the extraction of noun phrases considered concept candidates by applying linguistic and statistic approaches. When accessed in a virtual learning environment, the hierarchies became a differentiated resource for vocabulary teaching and for the journalistic language patterns research in newspapers geared to audiences with lower education.*

Resumo. *Neste artigo são apresentadas hierarquias de conceitos geradas automaticamente a partir do jornal popular Diário Gaúcho. As hierarquias são obtidas pela ferramenta ExATOlp através da extração de sintagmas nominais considerados candidatos a conceitos mediante combinação de técnicas linguísticas e estatísticas. Acessadas em um ambiente virtual de aprendizagem, as hierarquias oferecem recurso diferenciado para o ensino de vocabulário e pesquisa sobre padrões da linguagem jornalística voltada para públicos de menor escolaridade.*

1. Introdução

Apresenta-se aqui um modelo de hierarquia automática de conceitos produzida a partir de uma amostra de textos do jornal popular *Diário Gaúcho* (doravante DG). A hierarquia, em diferentes formatos, é gerada pela ferramenta ExATOlp, acessada gratuitamente em <http://www6.ufrgs.br/textecc/index_porpopular.php>, em que se oferece um ambiente virtual de aprendizagem. Seu acesso é um resultado parcial da pesquisa *Padrões do Português Popular Escrito – Projeto PorPopular*, dedicada a reconhecer o vocabulário e especificidades do texto de jornais populares brasileiros, cujos leitores preferenciais têm poder aquisitivo e graus de escolaridade menores se comparados aos de leitores de jornais tradicionais. A investigação tem apoio do CNPq e conta com a colaboração de pesquisadores do grupo de pesquisa de Processamento de Linguagem Natural (doravante PLN) da PUCRS <http://www.inf.pucrs.br/~linatural/>.

A ferramenta ExATOlp [Lopes *et al*. 2009] retira do *corpus* DG sintagmas nominais candidatos a conceitos (portadores de informação) via combinação de técnicas de extração baseadas em princípios de análise linguística e estatística. Para os itens extraídos, a ferramenta organiza uma hierarquia e uma lista dos contextos verbais nos quais cada conceito foi encontrado. A hierarquia gerada pode ser consultada em diferentes formatos, com maior ou menor detalhamento de informações, dependendo da opção de visualização do usuário, podendo auxiliá-lo na percepção da organização do conteúdo do *corpus* em foco.

Este trabalho está assim organizado: na Seção 2, está a caracterização do quadro geral da pesquisa PorPopular, do *corpus* e do jornal **DG**; na Seção 3, descrevem-se a ferramenta ExATOlp**,** o processo de geração das hierarquias e traz-se uma pequena amostra de hierarquias obtidas; na Seção 4, exemplificam-se aplicações das hierarquias, apresentam-se limitações do trabalho e são indicadas possibilidades para trabalhos futuros.

2. A Pesquisa PorPopular e o *Corpus* Reunido

A pesquisa PorPopular é de natureza linguística, centrada em aspectos lexicológicos e discursivo-gramaticais. Adota a metodologia e ponto de vista da Linguística de *Corpus* (doravante LC) [Berber Sardinha 2004] e enfatiza a descrição com base estatística partindo de acervos textuais em formato digital. Esses acervos são denominados *corpus*/*corpora* e servem tanto para estudos da linguagem quanto para produção de alguma aplicação computacional. A LC concebe a língua como um sistema probabilístico de combinatórias, de modo que não se pode observar as palavras isoladas do vocabulário de um texto ou *corpus*. Isso porque, conforme Stubbs [2001], o conhecimento humano da linguagem e dos textos não se restringe a um conhecimento das palavras isoladas, mas é integrado fundamentalmente pelo conhecimento de combinatórias possíveis e pelo conhecimento cultural que essas combinatórias frequentemente contêm. Assim, os principais focos da pesquisa PorPopular são a descrição e o estudo de padrões associativos do vocabulário para o que são utilizados também métodos, abordagens e produtos do PLN. Visa-se uma caracterização do léxico e da linguagem posta em um texto que é feito, em tese, de um modo mais simplificado, para ser compreendido com facilidade por pessoas com grau de escolaridade relativamente baixo. Na etapa atual da investigação, até o final de 2011, utiliza-se como *corpus* apenas textos coletados do jornal DG, versão impressa, publicado em Porto Alegre-RS, produzido pelo grupo RBS.

O DG impresso não oferece assinatura e é vendido apenas em bancas da cidade de Porto Alegre e região metropolitana. Foi escolhido para estudo em função de sua grande tiragem (168 mil exemplares/dia) e de sua longa existência (11 anos), além de já ter sido objeto de pesquisas na área do Jornalismo [Amaral 2006; Bernardes 2004]. Entretanto, ainda não havia sido explorado no âmbito dos Estudos da Linguagem ou do PLN. Seu número de leitores supera o de jornais da mesma cidade dirigidos a públicos mais tradicionais distribuídos em todo o Estado do Rio Grande do Sul. No *corpus*, estão arquivos de edições completas, coletadas em dias alternados da semana, do jornal impresso em formato somente texto (.txt) do ano de 2008, com pequenas amostras de 2009 e de 2010. O material em formato .txt pode ser compartilhado com pesquisadores e boa parte já se acessa via expressões de busca na seção **Experimente** do *site* do Projeto PorPopular. Materiais e recursos associados a esse *corpus* já estão sendo utilizados para atividades de ensino de língua portuguesa, ensino de vocabulário, como também integram proposta de um dicionário de português como língua estrangeira, dado um caráter *a priori* mais simples dos textos e da linguagem do DG. Entre as aplicações disponíveis, destaca-se neste trabalho o recurso **Hierarquias de Conceitos**, compreendido como uma representação ontológica do conteúdo dos textos reunidos, conforme detalhado na Seção 3 a seguir.

3. A Ferramenta ExATOlp e o Processo de Geração de Hierarquias

O processo de aquisição de hierarquias da ferramenta ExATOlp [Lopes *et al*. 2009] possui base linguística, uma vez que o ponto de partida é a extração de termos (sintagmas nominais), partindo de um *corpus* previamente anotado pelo *parser* PALAVRAS [Bick 2000]. Os sintagmas extraídos passam por uma análise detalhada em que diversas heurísticas são utilizadas para descartar ou aprimorar a qualidade informacional dos sintagmas obtidos.

Em resumo, o processo de extração de termos aplicado consiste em considerar todos os termos (*multi-token*) anotados como sintagmas nominais, ou *tokens* anotados pelo PALAVRAS como sujeito, objeto, complemento de sujeito ou objeto de orações. Em seguida, aplica-se um conjunto de heurísticas propostas na

ferramenta ExATOlp**.** A aplicação dessas heurísticas foi comparada com uma abordagem estatística em [Lopes *et al*. 2010] e o resultado da comparação mostrou a superioridade de cerca de 15% de precisão do método linguístico frente à abordagem puramente estatística. As heurísticas são as seguintes:

- ✓recusar termos que possuem numerais na sua forma textual ou numérica, por exemplo, "sete meses"; "8 horas", *etc*.;
- ✓recusar termos que tenham outros caracteres além de letras acentuadas ou não, por exemplo, "%", "\", "/", "@", *etc*.;
- ✓recusar termos que tenham como núcleo palavras identificadas sintaticamente com outras categorias além de substantivos comuns, próprios, adjetivos ou verbos no particípio passado, por exemplo: "**Eles mesmos** têm de construir um muro". Nesse caso, esse sintagma seria recusado, pois o núcleo é um pronome;
- ✓remover artigos contidos nos termos, por exemplo, "**O Incrível Hulk**" será salvo como "**Incrível Hulk**";
- ✓remover pronomes contidos nos termos, por exemplo, "Ele foi para **sua casa de praia**", o sintagma salvo será **casa de praia**;
- ✓criar termos implícitos detectados pelo uso de conjunções entre adjetivos, por exemplo, "As **pessoas espertas** ou sábias...", nesse caso, dois termos serão salvos: "**pessoas espertas**" e "**pessoas sábias**";
- ✓criar termos genéricos pela remoção sucessiva de adjetivos, por exemplo, do sintagma "**O perigo das doenças virais hemorrágicas**", cria-se mais três termos "**perigo das doenças virais**", "**perigo das doenças**" e "**perigo**"; e
- ✓replicar termos que são sujeitos de mais do que um predicado, por exemplo, "**Pacientes idosos** compram e tomam remédios caros" replica os termos, desdobrando a frase em duas: "**Pacientes idosos** compram **remédios caros**" e "**Pacientes idosos** tomam **remédios caros**".

Os termos considerados segundo a aplicação das heurísticas são salvos em listas de acordo com o número de palavras que os compõem, ou seja, unigramas (uma palavra), bigramas (duas), trigramas (três), *etc*. Além da anotação sintática, o PALAVRAS realiza anotação semântica dos *tokens* de acordo com um conjunto de 16 categorias. Com base nessas etiquetas semânticas, os termos são organizados em uma estrutura de árvore hiperbólica, apresentada na Figura 1.

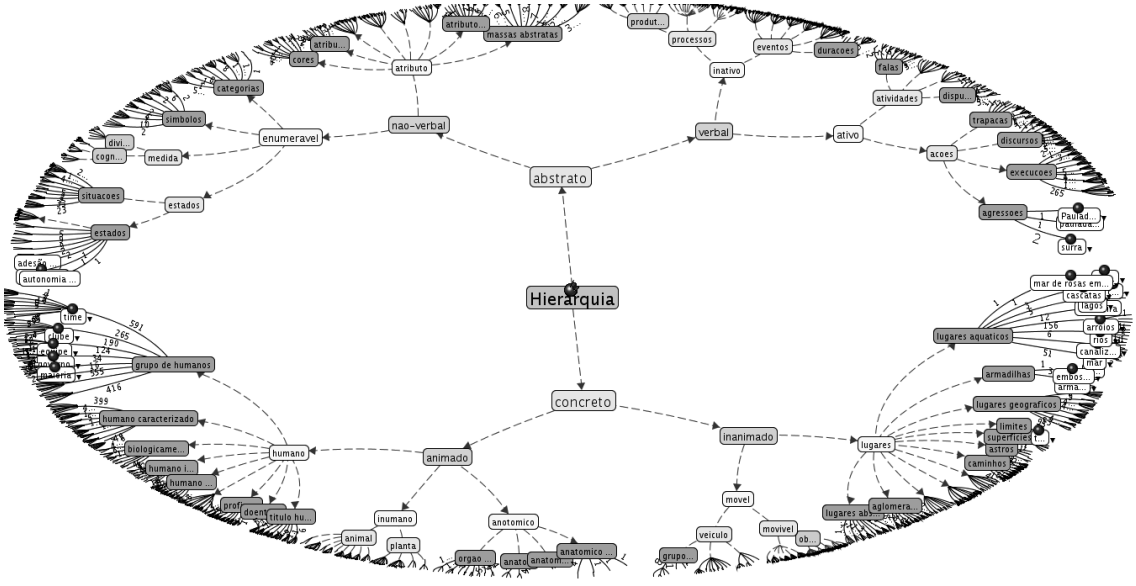


Figura 1. Visão geral do primeiro nível (categorias semânticas) da Hierarquia

Cada uma das categorias semânticas possui diversas etiquetas semânticas, totalizando 174 etiquetas distintas. Dessa forma, cada um dos termos extraídos (que correspondem às folhas da árvore) é relacionado a uma dessas etiquetas. A hierarquia baseada em categorias e etiquetas semânticas produz um primeiro nível qualificado. Em seguida, é feita uma nova estruturação a partir de cada etiqueta semântica, organizando o segundo nível da hierarquia. Nesse segundo nível, os termos (sintagmas nominais) são divididos em subgrupos que possuem o mesmo núcleo. Em seguida, é feita uma organização dos termos de cada subgrupo, considerando hiperônimo de um termo o termo que estiver contido nele.

Do *corpus* desta pesquisa, que possui 1.137.847 palavras, foram extraídos inicialmente 176.287 sintagmas nominais. Após aplicação das heurísticas de refinamento, 83.107 foram rejeitados. Os sintagmas aproveitados foram organizados em uma hierarquia em dois níveis: no primeiro, o nível das categorias semânticas; e, no segundo, nível hierarquia estrutural do termo. Por exemplo, o nodo **mulher** liga-se ao nodo **humano** num primeiro nível. Dessa forma, coloca-se hierarquicamente em relação a outros nodos semanticamente próximos, como **homem**, por exemplo. Partindo da estrutura lexical, são considerados filhos do nodo **mulher** os nodos que correspondem aos sintagmas **mulher de minha vida – mulheres nascidas este dia – mulheres que vivem em Rio de Janeiro – mulheres escravas**, etc.

A saída de todo o processo de geração de hierarquias realizado pela ferramenta ExATOlp são arquivos no formato XML. Arquivos desse tipo podem ser utilizados como formato de entrada de vários sistemas com diferentes aplicações. Uma delas é a visualização hiperbólica da hierarquia para analistas do discurso/texto, como nesta pesquisa com o DG. Nessa modalidade, é possível a manipulação dos termos por árvores interativas que facilitam tanto uma visão ampla dos nodos da hierarquia (Figura 2) quanto uma visão mais restritiva somente do segundo nível (Figura 3). Adicionalmente, as hierarquias possuem um terceiro nível, que corresponde aos verbos associados aos nodos, indicando-se a ocorrência do termo e seu respectivo contexto verbal. A Figura 4 traz a hierarquização do termo **arroios** e seus respectivos verbos. A hierarquia estabelece a seguinte ordem: hierarquia > concreto > inanimado > lugares > lugares aquáticos > **arroios** (que está relacionado com os verbos: **limpar**, **alagar**, **ter**, **transbordar**, **encontrar**, **apresentar**, **ser**). Note-se que o termo **arroios** possui hipônimos (**arroio local**, **arroio que corta a restinga**, *etc*.) que só aparecem quando não se observa seus contextos verbais (Figura 3).

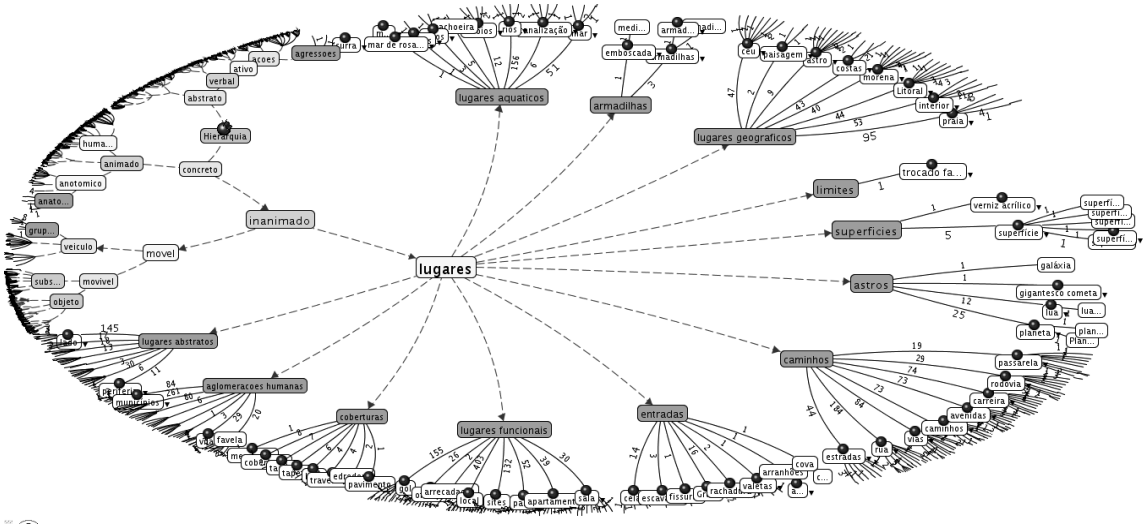


Figura 2. Visão focada na etiqueta semântica <lugares> para o *corpus* DG

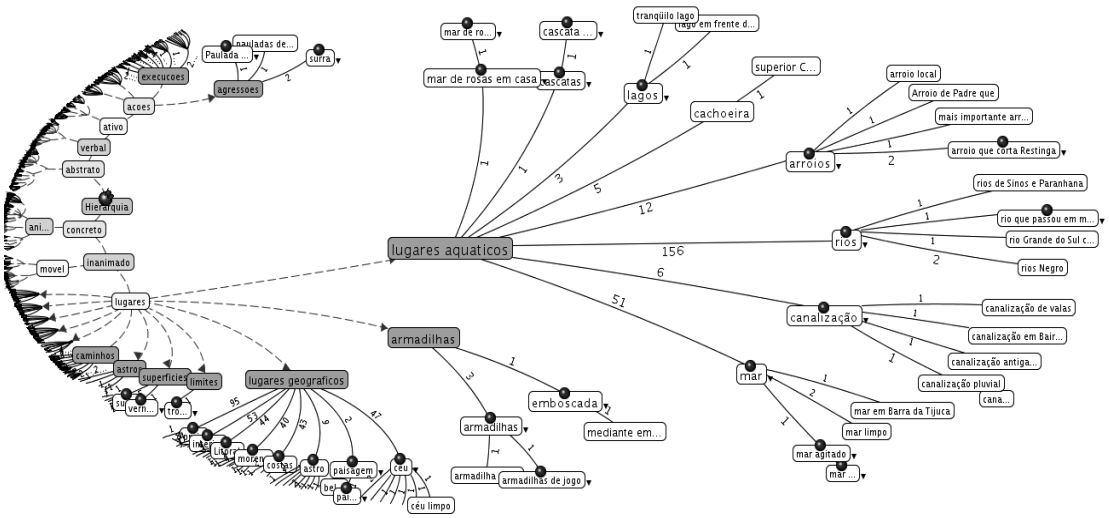


Figura 3. Visão do segundo nível da Hierarquia para <lugares> no *corpus* DG

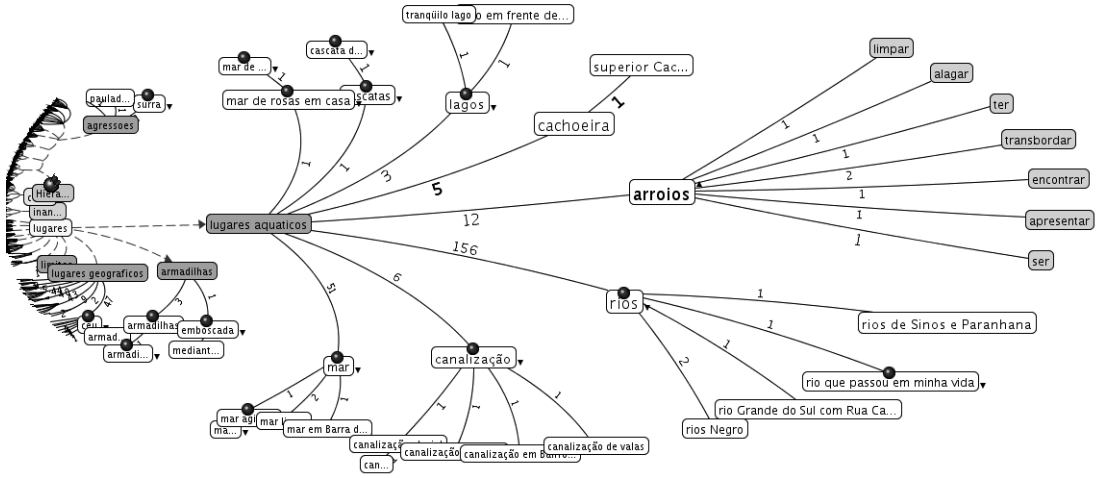


Figura 4. Visualização dos verbos na Hierarquia de conceitos do *corpus* DG

4. Utilizações das Hierarquias Geradas, Limitações e Trabalhos Futuros

A aplicação mais direta das hierarquias produzidas pela ferramenta ExATOlp é a geração de dados para auxiliar a reconhecer padrões de organização de conteúdo de um *corpus*, o que permite também um aproveitamento para o reconhecimento de padrões vocabulares e textuais em diferentes cenários comunicativos e em distintas categorias de textos (redações escolares, textos jornalísticos, textos científicos, *etc*.) que conformem um dado *corpus*. No caso do jornalismo popular, um segmento ainda pouco estudado entre os pesquisadores de Comunicação Social/Jornalismo e de Letras/Linguística, o auxílio é bastante importante, sobretudo pelo tipo e desenho de informação que o recurso oferece ao investigador da linguagem em foco.

Como uma outra aplicabilidade futura associada também ao *corpus* DG, embora indireta, cita-se a construção de um dicionário *on-line* de português para estrangeiros, projeto em fase inicial ao Projeto TEXTECC <http://www6.ufrgs.br/letras/dicionarioportuguesle/>. Com novas hierarquias geradas desse

corpus DG, segmentado por temáticas ou assuntos do jornal, esse dicionário *on-line* poderia dispor seus verbetes por nodos em vez de privilegiar uma ordem estritamente alfabética, ou até mesmo oferecer definições e verbetes dinâmicos, permitindo visualização da rede de relações estabelecida através da saída da ferramenta. Por meio da apresentação da hierarquia de conceitos, o usuário teria uma visão amplificada, por exemplo, de um determinado domínio semântico e que conseguiria estabelecer ou reconstruir as relações de forma mais dinâmica e proveitosa, construindo seu conhecimento de língua de forma mais concreta. Esse recurso também teria bom aproveitamento em dicionários para falantes nativos do português.

O trabalho realizado também poderia ser expandido para que as hierarquias incluam outras relações além da relação taxonômica "é um" representada na árvore hiperbólica. Assim, seria possível localizar termos que são sujeitos e objetos de verbos específicos e criar relações não-taxonômicas. Uma opção, por exemplo, seria extrair da ocorrência de uma frase tal como "Os alunos compram doces" e criar uma relação "comprar" entre o termo "aluno" e o termo "doce". Este tipo de relação não-taxonômica seria fácil de extrair a partir da versão atual da ferramenta, mas seria necessário encontrar uma forma de escolher os verbos para criar as relações e o modo de visualização desse tipo de informação, pois uma árvore hiperbólica não se prestaria para isso, já que o conjunto de relações na maioria dos casos poderia gerar ciclos. Cabe salientar que o processo de aquisição de hierarquia de conceitos executado automaticamente pela ferramenta integra um trabalho de doutorado em andamento junto ao Grupo de Processamento da Linguagem Natural da Faculdade de Informática da PUCRS, de modo que novas configurações para a seleção dos sintagmas podem ser testadas.

A ferramenta ExATOlp gera representações de conteúdo dos textos que se prestam a um sem-número de aplicações. Integrada a um ambiente virtual de aprendizagem que aproveita textos de um jornal popular, oferece uma visualização panorâmica sobre seu conteúdo em diferentes opções. Da integração entre o PLN, a geração de ontologias e os Estudos da Linguagem, especialmente com a LC, tem-se uma nova opção de leitura e de descobertas para os conteúdos e palavras postos nesse tipo de texto, sem contar que a amplitude e a dinamicidade das representações de conteúdo de textos e dos *corpora* tende a entusiasmar estudantes e pesquisadores.

Referências

Amaral, M. F. (2006), Jornalismo Popular, São Paulo, Contexto.

Berber Sardinha, T. (2004), Linguística de Corpus, Barueri, São Paulo, Manole.

Bernardes, C. B. (2004), As Condições de produção do jornalismo popular massivo: o caso do Diário Gaúcho, Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-Graduação em Comunicação e Informação, Diss. Mestrado.

Bick, E. (2000), The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, PhD thesis, Arhus University.

Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. (2009) ExATOlp – "An automatic tool for term extraction from Portuguese language corpora" In *Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a challenge for computer science and linguistics (LTC'09)*. Adam Mickiewicz University.

Lopes, L.; Oliveira, L. H.; Vieira, R. (2010) "Portuguese term extraction methods: comparing linguistic and statistical approaches" In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*.

Stubbs, M. (2001), Words and phrases: Corpus studies of lexical semantics, Oxford, Blackwell.

Interoperabilidade e portabilidade de documentos digitais usando ontologias

Erika Guetti Suca [1], **Flávio Soares Corrêa da Silva**[1]

[1]Instituto de Matemática e Estatística - Universidade São Paulo (IME-USP)
São Paulo – SP – Brasil

{eguetti,fcs}@ime.usp.br

Abstract. *Our purpose is to enable* ***interoperability of documents*** *and achieve* ***portability*** *of* ***digital documents*** *through the reuse of content and format in different plausible combinations. We propose the characterization of digital documents using* ***ontologies*** *as a solution to the problem of lack of interoperability in the implementations of document formats. As proof of concept we consider the portability between* ***ODF (Open Document Format)*** *and* ***OOXML (Office Open XML)*** *document formats.*

Resumo. *Nosso objetivo é possibilitar a* ***interoperabilidade de documentos*** *e atingir a* ***portabilidade*** *simples e confiável de* ***documentos digitais*** *através da reutilização de formatos e conteúdos, em diferentes combinações plausíveis. Propomos a caracterização de documentos digitais usando* ***ontologias*** *como solução ao problema da falta de interoperabilidade nas implementações de formatos de documentos. Como prova de conceito, será considerada a portabilidade entre os formatos de documentos* ***ODF (Open Document Format)*** *e* ***OOXML (Office Open XML)****.*

1. Introdução

As organizações precisam trocar informação através de documentos. Muitas vezes esses documentos são apresentados com formato e conteúdos pré-definidos, que podem ser equivalentes ou quase equivalentes entre si, porém bastantes distintos em diferentes organizações (ou em uma mesma organização em diferentes contextos históricos). Como recurso importante para gerenciar seu conhecimento de forma efetiva e preservar seu capital intelectual, as organizações precisam disponibilizar documentos independentemente do *software* com que foram criados. Propomos a caracterização dos formatos de documentos digitais usando ontologias para favorecer a portabilidade e superar o problema de falta de interoperabilidade de documentos nas organizações.

O trabalho está organizado da seguinte forma: na Seção 2 introduzimos o problema da preservação dos documentos digitais; na Seção 3 apresentamos os conceitos fundamentais da interoperabilidade e portabilidade de documentos; na Seção 4 mostramos os principais conceitos das ontologias; na Seção 5 resumimos alguns trabalhos relacionados. Na Seção 6 explicamos nossa proposta; finalmente, na Seção 7 fazemos as considerações finais.

2. Preservação dos documentos digitais

Um **documento digital** é um documento codificado em formato binário, acessível por meio de um sistema computacional [Gouget et al. 2005]. Nosso trabalho está focado em

documentos digitais criados a partir de aplicações de escritório. As aplicações de escritório são aplicativos voltados para as tarefas de escritório e geralmente estão agrupadas em interfaces de usuário conhecidas como suítes de escritório.

Cada documento digital possui um formato de arquivo. O **formato de arquivo** especifica a estrutura em que os códigos digitais estão organizados [Shepard and MacCarn 2008]. A codificação do formato de arquivo está profundamente relacionada ao programa que o criou. Às vezes, após um período de tempo, se torna extremamente difícil a leitura do documento sem perda significativa da informação. Assim, os documentos podem existir por dezenas de anos, mas a vida útil de uma suíte de escritório não é sempre garantida. Diante disso, as organizações têm que garantir a integridade e perpetuidade dos documentos, mesmo após o *software* que os criou ter desaparecido do mercado [Ngo 2008, Taurion 2009].

Na preservação de documentos é importante provar sua autenticidade. A **autenticidade** é a capacidade de demostrar que um documento digital é aquilo que se propõe ser. É fundamental provar que existe um conjunto de propriedades significativas que foram corretamente preservadas ao longo do tempo. Quanto maior for o número dessas propriedades, maiores serão os requisitos relativos à infraestrutura tecnológica para dar suporte à sua preservação. Torna-se necessária a criação de políticas de preservação que exprimam, para cada classe de objetos digitais, o conjunto de propriedades significativas que serão asseguradas pelo repositório [Ferreira 2006, Rusbridge 2003].

Este trabalho concentra-se na preservação da autenticidade dos documento digitais, mais do que na preservação das características estéticas do documento.

3. Interoperabilidade e portabilidade de documentos

Interoperabilidade[1] é a habilidade de transferir e utilizar informações de maneira uniforme e eficiente entre várias organizações e sistemas de informação. A **interoperabilidade de documentos** é a habilidade das aplicações de documentos de extrair dados de diferentes tipos de documentos e transformá-los em estruturas padronizadas. Esta informação pode ser trocada entre vários sistemas e posteriormente ser processada [Schmidt et al. 2006]. Com a interoperabilidade de documentos baseada principalmente na tecnologia XML, as aplicações podem comunicar-se diretamente com serviços do governo eletrônico.

Por outro lado, a **portabilidade de documentos** é a troca de documentos com todas as informações que eles contêm, principalmente suas configurações de formato [Schmidt et al. 2006]. Na portabilidade de documentos é importante a fidelidade do formato do documento. A **fidelidade de formato** é a capacidade de manter o formato do documento e seu sentido associado, apesar de ser editado em múltiplas aplicações [Ditch 2007].

Ao contrário da portabilidade, a interoperabilidade de documentos está exclusivamente preocupada com a troca da dados corporativos contidos nos documentos e não faz exigências sobre requisitos em termos de aparência, elementos estilísticos, formato ou questões semelhantes [Schmidt et al. 2006].

[1] http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade/o-que-e-interoperabilidade

3.1. *Office Open XML* (OOXML) e *OpenDocument Format* (ODF)

OOXML e ODF são os principais padrões abertos de formatos de documentos baseados em XML. No entanto, os dois padrões são incompatíveis e rivais no mercado, provocando a **guerra dos formatos abertos**, gerando discussões técnicas sobre as vantagens e desvantagens de cada um deles.

O **OOXML** foi desenvolvido pela *Microsoft* em 2008 e tornou-se um padrão aberto ISO (ISO/IEC 29500:2008). OOXML foi projetado para representar o corpus preexistente de documentos de processamento de texto, apresentações e planilhas que são codificados pela *Microsoft*. A especificação OOXML contém material normativo e informativo estruturado em aproximadamente 6546 páginas.

O **ODF** foi desenvolvido pela Sun Microsystems em 2002 e seu processo de padronização foi iniciado pela OASIS[2]. O ODF foi projetado para ser uma especificação de formato de documentos independente de fornecedor ou de *software*. Embora a especificação ODF (aproximadamente 700 páginas) seja complexa para os padrões normais, a reutilização de padrões abertos existentes reduz consideravelmente a complexidade da especificação [Eckert et al. 2009, Ngo 2008].

O estudo de [Shah and Kesan 2009] demonstrou que não existem implementações que ofereçam 100% de compatibilidade (avaliação da leitura dos documentos) dentro das implementações de OOXML e ODF.

4. Ontologias

Na ciência da computação, a definição mais citada de ontologia na literatura é de [Gruber 1993]: uma **ontologia** é uma especificação explícita de uma conceitualização. Em 1997, Borst [Borst 1997] ligeiramente modifica a definição de Gruber, afirmando que: uma **ontologia** é uma especificação formal de uma conceitualização compartilhada. A formalização de uma ontologia está definida em cinco componentes: conceitos, relações, funções, axiomas e instâncias [Gruber 1993]. As ontologias, dependendo do seu grau de formalidade, podem ser modeladas baseadas em técnicas de modelagem de inteligência artificial, engenharia de *software* e bancos de dados [Gómez-Pérez et al. 2005].

A ontologia explicita a informação independentemente das estruturas de dados que são usadas para armazenar a informação. Além disso, a conceitualização de uma ontologia pode ser expressa em várias linguagens [Guimarães 2008]. Elas são projetadas para que a informação seja compartilhada entre agentes que garantam compromissos ontológicos. Desse modo, as ontologias viabilizam soluções para problemas como a falta de padronização, falta de interoperabilidade, problemas com a recuperação da informação, falta de reuso, confusões terminológicas, problemas de troca de informações entre agentes de *software*, dentre muitos outros [Uschold and Gruninger 1996].

Este trabalho propõe o uso das ontologias para a modelagem das características da estrutura de formato e do conteúdo corporativo dos documentos digitais.

5. Trabalhos Relacionados

No trabalho de [Eckert et al. 2009] é analisado como os formatos OOXML e ODF especificam as características mais importantes dos documentos e como essas características

[2]http://www.oasis-open.org

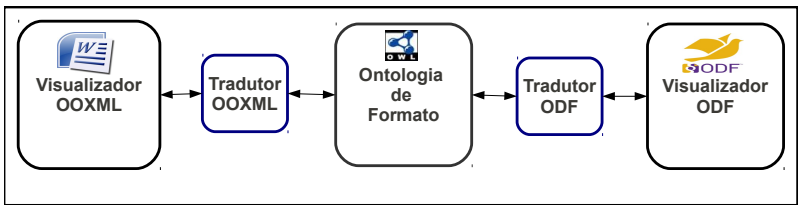


Figura 1. Ontologia como interlíngua.

podem ser traduzidas entre os dois formatos. Esse trabalho concentra-se mais na definição de orientações para a tradução da estrutura da apresentação do documento do que sobre a preservação da estética do documento. Os autores concluíram que a separação da apresentação do documento do seu conteúdo corporativo oferece mais facilidade na manipulação de documentos. Assim, editar os componentes de apresentação e dados do conteúdo corporativo de forma independente confere flexibilidade considerável na criação e edição de documentos.

Outro trabalho é de [Eriksson 2007]. O trabalho dele explica como a combinação de ontologias e documentos cria novas possibilidades para melhorar a gestão do conhecimento nas organizações, isso através dos documentos semânticos. Os documentos semânticos são as integrações dos documentos com as ontologias. O trabalho de [Eriksson 2007] integrou documentos no formato PDF com três ontologias: ontologia de anotação, ontologia do documento e uma ontologia do domínio. Essas ontologias ajudam na explicação do conteúdo do documento e facilitam sua busca. As múltiplas ontologias permitiram ter conceitualizações com diferentes intenções habilitando seu reúso.

6. Caracterização de documentos digitais usando ontologias

Propomos a construção de um modelo que considere as qualidades essenciais de formato de um documento digital. Um documento digital criado com o padrão OOXML ou ODF poderá ser caracterizado nesse modelo.

Nosso modelo será caracterizado a partir de ontologias. Uma **ontologia do formato** especializada em caracterizar a estrutura da apresentação do documento (parágrafos, tabelas, listas, enumerações, etc.), ela apresenta a informação da ontologia de conteúdo. Outra **ontologia de conteúdo** que caracteriza a informação dos dados corporativos contidos no documento (dados do paciente, estudante, vendas, etc.).

A Figura 1 ilustra a interação da ontologia de formato e o documento final. Os documentos digitais serão recriados através de tradutores. Os tradutores são mediadores entre a ontologia de formato e um formato especifico.

A ontologia de formato é criada baseada em um conjunto de propriedades significativas a serem preservadas de uma classe específica de documentos. O objetivo da criação da ontologia de formato é atingir a portabilidade simples do documento, isto é, preservar a estrutura da organização da apresentação do documento, a Figura 2 ilustra os conceitos principais desta ontologia. Por outro lado, o objetivo da criação da ontologia de conteúdo é possibilitar a interoperabilidade de documentos, isto é, habilitar o intercâmbio coerente de informações corporativas específicas sobre um contexto.

A ontologia de formato pode ser modificada independentemente da ontologia de

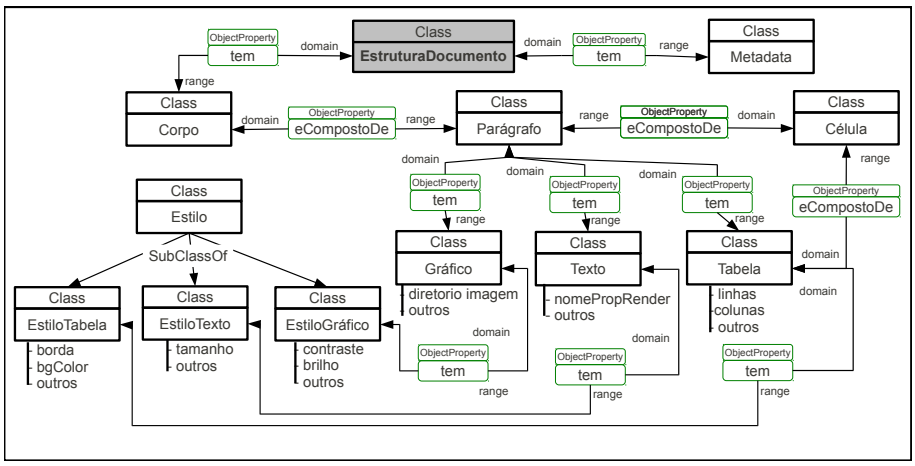


Figura 2. Ontologia de Formato.

conteúdo e vice-versa. Os documentos digitais serão recriados através de tradutores entre a ontologia de formato e um formato específico. Um novo documento pode ser construído com a mesma informação contextual, mas agora num formato apropriado a seus propósitos, bem como um novo documento pode ter a mesma informação contextual, mas com uma apresentação distinta.

6.1. Avaliação dos resultados

No trabalho de [Shah and Kesan 2009] é apresentado um exemplo de avaliação de interoperabilidade e portabilidade de documentos. Eles testaram um conjunto de suítes de escritório que implementam os formatos de documentos ODF e OOXML. Os resultados do estudo estão baseados em pontuações de quão bem as implementações podem ler e escrever documentos. A parte final da pontuação está focado na capacidade de preservação dos metadados dos documentos, isto é, atributos de estilos, números de páginas, tabelas de conteúdos ou cabeçalhos, informações do documento (tempo, ou o número de palavras em documentos), e controle de alterações.

A avaliação dos resultados do nosso trabalho seguirá essa metodologia de avaliação, no entanto com adaptações. Neste caso os documentos serão traduzidos para diferentes formatos, julgando que cada documento pode ser traduzido seguindo vários graus de fidelidade. Para cada documento recriado a partir da ontologia de formato, será quantificada qualquer modificação ao conteúdo original.

7. Considerações finais

A interoperabilidade é um ponto crítico nas questões de governo eletrônico. O presente artigo apresenta a problemática na preservação de documentos digitais e destaca a importância da interoperabilidade e portabilidade nos formatos de documentos digitais. Propomos o uso de ontologias para garantir a integridade e perpetuidade dos documentos digitais, priorizando a preservação das características da estrutura da apresentação e conteúdo corporativo (os elementos importantes para provar sua autenticidade) sobre a preservação das características de estética do documento.

Como trabalho futuro será desenvolvido um caso de uso de interoperabilidade de documentos aplicada ao governo eletrônico, em que as ontologias de formato e conteúdo

tenham papel comprovado na preservação e distribuição eficientes de documentos digitais.

Referências

Borst, W. (1997). *Construction of Engineering Ontologies*. PhD thesis, University of Tweenty, Enschede, NL, Centre for Telematica and Information Technology.

Ditch, W. (2007). *XML-based Office Document Standars*. JISC: Bristol, UK, United Kingdom.

Eckert, K.-P., Ziesing, J., and Ishionwu, U. (2009). *Document Operabilitiy Open Document Format and Office Open XML*. Fraunhofer Verlag, Germany, fokus edition.

Eriksson, H. (2007). The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, 65:624–639.

Ferreira, M. (2006). *Introdução e preservação digital: Conceitos, estratégias e actuais consensos*. PhD thesis, Escola de Engenharia da Universidade do Minho.

Gouget, A. G., Monteiro, B. M., Santos, C. R., da Silva Maçulo, E., de Oliveira, M. I., Miguel, M. L. C., Sobrosa, N. B. S., de Moura Estevão (coord.), S. N., de Mello Lopes, V. L. H., and da Fonseca, V. M. M. (2005). *Dicionario Brasileiro de Terminologia Arquivística*. Arquivo Nacional (Brasil), Rio Janeiro. Edição e Revisão Alba Gisele Gouget and Silvia Ninita de Moura Estevão and Vera Lucia Hess de Mello Lopes and Vitor Manoel Marques da Fonseca.

Gruber, T. R. (1993). A translation approach to portable ontology specification. In *Knowlegde Acquisition*, pages 199–220.

Guimarães, F. J. Z. (2008). Utilização de ontologias no domínio b2c.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2005). *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag, 4 edition.

Ngo, T. (2008). Visão geral do office open xml. Technical report, Ecma International.

Rusbridge, A. (2003). Migration on request. *4th Year Project Report Computer Science*.

Schmidt, K.-U., Fox, O., Henckel, L., Holzmann-Kaiser, U., Martin, P., and Tschichholz, M. (2006). *Document Interoperability for Use in eGovernment. Integration of XML-based Document Content in Public Administration Processes*. Fraunhofer Verlag, fokus edition.

Shah, R. and Kesan, J. (2009). Interoperability challenge for open standars: Odf and ooxml as examples. *The proceedings of the 10th International Digital Government Research Conference*.

Shepard, T. and MacCarn, D. (2008). The universal preservation format: A recommended practice for archiving media and electronic records.

Taurion, C. (2009). Adoptando o odf como padrão aberto de documento. Technical report, ODF Alliance Brasil. Volume 1.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, 11.

Ontologias no Suporte a Evolução de Conteúdos em Portais Semânticos

Débora Alvernaz Corrêa[1], Maria Cláudia Cavalcanti[1], Ana Maria de C. Moura[2]

[1]Departamento de Sistemas e Computação
Instituto Militar de Engenharia (IME) - Rio de Janeiro, RJ – Brasil
[2]Extreme Data Lab (DEXL Lab)
Laboratório Nacional de Computação Científica (LNCC) - Petrópolis – RJ – Brasil
{deboradac,anamaria.moura}@gmail.com, yoko@ime.eb.br

Abstract. *In a semantic portal, contents are described and organized based on domain ontologies. However, with the increasing amount of information generated each day on the web, dynamic publishing in these portals, whose contents are obtained from a large and diversified number of sites, still represents a major challenge, since this task lacks mechanisms to update and integrate information automatically. This paper presents an architecture that facilitates the population of a domain ontology from web sites that have a certain semantic feature. These instances may be used later in the process of a semantic portal automatic updating.*

Resumo. *Em um portal semântico, conteúdos são descritos e organizados com base em ontologias de domínio. Entretanto, com a quantidade crescente de informações geradas a cada dia na web, a publicação dinâmica nesses portais, cujos conteúdos são oriundos de um grande e diversificado número de sites, ainda representa um grande desafio, uma vez que essa tarefa carece de mecanismos para atualizar e integrar informações automaticamente. Este artigo apresenta uma arquitetura que facilita a população de uma ontologia de domínio a partir de sites que apresentam alguma característica semântica. As instâncias recuperadas destes sites podem ser utilizadas posteriormente no processo de atualização automática de um portal semântico.*

1. Introdução

A Web Semântica surgiu com a finalidade de suprir as deficiências da web atual. De acordo com [Berners-Lee et al., 2001], significa disponibilizar informações com significados adicionais, de forma a contextualizá-los e torná-los interpretáveis por máquina, permitindo que agentes e pessoas possam trabalhar em cooperação.

Neste contexto surgiram os portais semânticos que, ao contrário dos portais tradicionais, agregam valores semânticos que ajudam na classificação e organização dos seus conteúdos, facilitando os mecanismos de busca a recuperarem informações mais úteis ao usuário final, isto é, com maior precisão. Os portais semânticos utilizam-se de ontologias [Gruber, 1995] como mecanismo básico para fornecer expressividade semântica a seu conteúdo. A tendência atual é adicionar às suas funcionalidades a capacidade de realizar consultas aos conteúdos da ontologia que embasam o portal via *endpoints*, utilizando a linguagem SPARQL[1]. No entanto, para alimentar portais semânticos é preciso buscar formas automatizadas de modo a mantê-los sempre atualizados. Muitos ainda contam com mecanismos manuais, como formulários.

[1] http://www.w3.org/TR/rdf-sparql-query/

Em [Latchim, 2008], é proposta uma arquitetura para recuperar informações da web baseadas em ontologias de domínio. Com essas informações recuperadas, é possível integrar o conteúdo do portal com as informações obtidas a partir de diferentes portais semânticos. No entanto, já que boa parte da informação que se deseja recuperar encontra-se em portais tradicionais ou simplesmente na web aberta, voltamos a enfrentar os problemas e limitações já conhecidos. Cada página web tem sua estrutura própria, textos mal formatados e carentes de metadados que possam descrevê-los, não havendo uma separação nítida entre o conteúdo e a apresentação da informação, o que interfere sobremaneira na qualidade dos serviços de busca.

Assim, o trabalho aqui proposto dá continuidade ao trabalho de [Latchim, 2008], tendo como objetivo a especificação de uma arquitetura que permita coletar conteúdos a partir de outros sites e/ou portais da web aberta, considerando um domínio específico, e instanciar à ontologia já existente que serve de base para um portal semântico, desse mesmo domínio, com tais conteúdos. Dessa forma, estaremos facilitando a alimentação deste portal semântico, e ajudando-o a manter-se atualizado.

O restante desse artigo está estruturado da seguinte forma. A seção 2 descreve alguns trabalhos relacionados. Na seção 3 é descrita a arquitetura proposta para alimentar portais semânticos através de sites e/ou portais web. A seção 4 apresenta um estudo de caso no qual, conteúdos de portais no contexto educacional, servem de subsídio para atualizar e popular uma ontologia nesse mesmo domínio. E por fim, a seção 5 conclui o artigo com alguns comentários e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Algumas alternativas de solução para o problema de interoperabilidade de informações entre portais e instanciação automática de conteúdo têm sido alvo de pesquisa há alguns anos. A literatura apresenta alguns trabalhos, tais como [Lachtim, 2008], [Lachtim et al., 2009], [Suominem et al., 2009], [Yvon et al., 2009] e [Castaño, 2008].

Esses trabalhos apresentam como característica comum o uso de tecnologias da WS, porém utilizados em contextos diferentes. Lachtim [Lachtim, 2008] integra e instancia informações a partir de portais semânticos. Em [Castaño, 2008] a população da ontologia é feita através de páginas HTML de currículos, mas com o objetivo de gerar relatórios. Já em [Suo minem et al., 2009] metadados e documentos são recuperados de conteúdos publicados nos Sistemas de Gerenciamento de Conteúdos (*Content Management Systems*) ou por conteúdos anotados manualmente, através do editor de metadados SAHA [Kurki et al., 2010]. Posteriormente, estes metadados são conectados aos serviços da ontologia ONKI [Viljanen et al., 2010] para serem validados. Os metadados validados com sucesso são então publicados no portal. O portal apresentado por [Hyvönen et al., 2009] utiliza como processo de criação de conteúdos uma variedade de esquemas de metadados, ferramentas como ONKI, SAHA, POKA [Poka, 2011] e VERA [Vera, 2011], além de serviços da Web 2.0, como *Wikipedia* e *Panoramio*. Esse processo permite a produção e recuperação de conteúdos relacionados a museus, bibliotecas, arquivos e outras organizações, cidadãos individuais e de fontes nacionais e internacionais da Web 2.0.

O grande diferencial do trabalho aqui proposto em relação aos demais está na atualização de portais semânticos a partir de conteúdos de sites e/ou portais da web aberta, considerando apenas a estrutura de apresentação e a estrutura navegacional dos portais, como *links*, listas e tabelas. Dessa forma, as atualizações destes portais são feitas de forma simples, permitindo assim que estes deixem de ser simples páginas voltadas somente para usuários, e tornem-se capazes de integrar e instanciar informações a partir do uso de ontologias.

3. Arquitetura Proposta para Alimentar Portais Semânticos

No escopo desse trabalho, o termo portal com potencial semântico é designado a todo aquele que se beneficie de uma das seguintes características: (i) tenha algum tipo de organização e hierarquia em sua estrutura; e/ou (ii) parte de seu conteúdo seja apresentado na forma de uma taxonomia, isto é, bastando que algumas de suas páginas apresentem tais características.

A figura 1 apresenta a arquitetura proposta por esse trabalho. É constituída de vários componentes que, em termos gerais, contribuem para alimentar uma ontologia a partir de portais da web aberta. A estratégia adotada por esta arquitetura segue os seguintes passos: a partir de uma navegação realizada em portais web com potencial semântico (lista de portais definida previamente pelo usuário), tendo como base uma ontologia de domínio (OB), informações consideradas úteis são extraídas para enriquecer esta ontologia com novas instâncias. Lembrando que este artigo não contempla a construção de uma nova ontologia, mas sim uma extensão da mesma. A seguir é apresentada uma nova versão da ontologia inicial, aqui denominada OB', com as novas informações ali encontradas em formato de triplas RDF[2]. Essa nova versão da ontologia OB' servirá como entrada para o alinhamento com as informações do portal semântico em questão. Dessa forma, as categorias existentes no portal semântico considerado serão atualizadas, tendo como base os novos conteúdos adicionados à OB'.

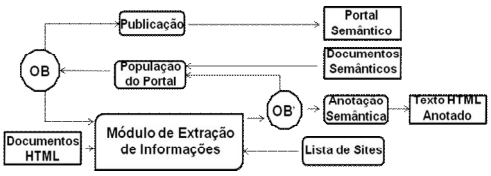


Figura 1. Arquitetura Proposta

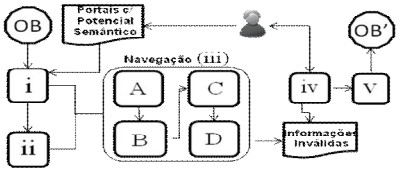


Figura 2. Módulo de Extração de Informações

O principal módulo dessa arquitetura diz respeito à Extração de Informações, que dará subsídios para a população de portais. Esse módulo, ilustrado na Figura 2, é composto por outros submódulos, cada um apresentando funcionalidades bem específicas, cujas características são descritas a seguir.

i. Recorte da OB. Esta etapa carrega uma lista de classes, instâncias e propriedades de relacionamentos existentes na Ontologia Base (OB) que servirão de base para a pesquisa de instâncias em cada página web visitada referente aos portais com potencial semântico. Os relacionamentos entre as classes são considerados para estabelecer a ordem de navegação pelas classes da ontologia. Além disso, é levado em consideração o nome real de cada classe (sempre começando pela classe mais abrangente definida pelo usuário), o *label* e as classes equivalentes para a navegação nas páginas (processo iii) dos portais. As instâncias de cada classe, bem como as instâncias equivalentes (definidas pela cláusula *same as*) às instâncias principais;

ii. Pré-categorização e identificação de página de início. Uma pré-categorização com base no título será efetuada para limitar a navegação estabelecida por (i). Caso a página inicial contenha no título um nome similar a uma instância de alguma classe da OB, a navegação se dará a partir da próxima classe a ser pesquisada. Caso contrário, a navegação se dará a partir da primeira classe. A identificação da página de início serve para definir a página a partir da qual será iniciada a navegação (processo iii). Se esta não for informada a navegação se inicia pela página principal do portal;

iii. Navegação pelas páginas. Este módulo realiza a navegação pelas páginas (de cada portal definido pelo usuário) a procura de *links*, que servirão para recuperar classes e instâncias. É

[2] RDF é uma linguagem ontológica. Triplas são declarações com a seguinte estrutura: "sujeito, predicado e objeto". Descreve a relação de um objeto a outro objeto ou literal através de um predicado. [RDF – W3C, 2011]

composto pelos subprocessos de *Recuperação de Classes, de Instâncias, de Pares de Instâncias* e *Análise de Hierarquia*, descritos a seguir.

A. *Recuperação de classes.* A navegação começa pela página de início pré-definida anteriormente pelo usuário no início do processo. O sistema deverá primeiramente ler a página a procura de *links* e *labels* que apresentem um grau de similaridade com a classe da OB procurada (definida no processo i). Estes *links* serão considerados prioritários para a navegação e deverão ser internos, i.e., do mesmo domínio de navegação. Quando um *link* que satisfizer as condições descritas anteriormente for encontrado, o sistema deverá verificar se este já foi visitado. Em caso afirmativo, irá para o próximo *link*, e em caso negativo, este deverá ser visitado e as instâncias recuperadas conforme o passo B;

B. *Recuperação de instâncias*. Para que instâncias sejam consideradas candidatas à alimentação de um portal semântico, estas deverão estar em *tags*[3] (*links*, listas e/ou tabelas). Além disso, devem ter o título com alguma similaridade (palavras semelhantes) com as instâncias existentes na OB (definida em i). Durante a navegação, as informações extraídas são recuperadas para posterior validação do usuário (processo iv), incluindo as triplas e os sites candidatos a portais com potencial semântico;

C. *Análise de hierarquia*. De modo a evitar a repetição de triplas na OB', as informações deverão estar compatíveis com a hierarquia definida pela OB. Por exemplo, uma instância de uma subclasse pode ser listada duas vezes, visto que é a mesma instância para superclasse;

D. *Recuperação de Relacionamentos*. Pares de instâncias em conformidade com os relacionamentos presentes na OB. Assim, por exemplo, na ontologia da Figura 3, as instâncias de "*Education_Program*" se relacionam com as de "*Academic_Research_Institution*" através da propriedade "*provided_By_Program*". Estes pares são então recuperados e armazenados como triplas RDF.

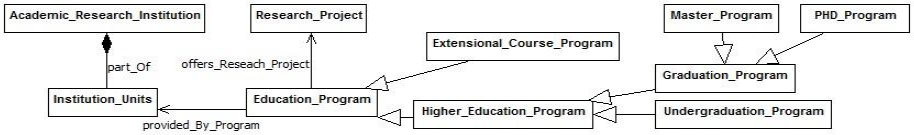


Figura 3. Recorte da OBEDU

iv. Validação. O usuário realiza a validação das informações extraídas. As que forem consideradas inválidas deverão ser armazenadas, para que posteriormente sirvam como uma pré-validação para as próximas informações recuperadas;

v. Transformação em triplas RDF. Este processo realiza a transformação das informações válidas em triplas RDF, que deverão ser adicionadas à nova ontologia, i.e., a OB'. Esta corresponde a um recorte vazio da OB (sem instâncias), que vai sendo atualizada com as novas instâncias recuperadas. Posteriormente a OB' deverá passar por um processo de alinhamento de ontologias com a OB, e suas instâncias poderão ser usadas para a população de um portal semântico cuja base seja a OB.

4. Estudo de Caso

A OB tem um papel muito importante na estratégia proposta para a alimentação de portais semânticos. Embora a arquitetura proposta tenha como foco uma aplicação genérica, essa seção apresenta um estudo de caso voltado para o domínio educacional, cuja ontologia base utilizada é

[3] *Tão* significa etiqueta e são utilizadas como breves instruções em linguagens de marcação.

a OBEDU [Lachtim et al., 2009], que serviu de ontologia base para a criação do portal semântico POSEDU[4].

A estratégia proposta na seção anterior foi adaptada para o contexto educacional. Nesta seção, foi utilizada apenas uma visão parcial da ontologia OBEDU, como mostrada na Figura 3. O recorte da OBEDU fornece classes e instâncias (recuperadas em i) para auxiliar a pesquisa nos portais com potencial semântico definidos pelo usuário. Inicialmente, o usuário realiza uma pré-categorização da página principal do portal (conforme mencionado anteriormente em ii), onde definirá a classe inicial de onde começará a navegação em busca de instâncias. As instâncias procuradas têm sempre como base as classes da ontologia. Assim, a página inicial é percorrida e seus *links* são extraídos começando pelos de maior prioridade em relação à OB (processo i). Os não prioritários deverão ser visitados posteriormente, até que as opções de extração de instâncias tenham sido esgotadas para uma determinada página. Um exemplo de navegação por um portal com potencial semântico é mostrado a seguir.

Neste exemplo, o portal do Instituto Militar de Engenharia - IME[5] foi escolhido como modelo para o estudo de caso (Figura 4). Na pré-categorização (processo ii), é verificado se o título da página principal do portal é similar a alguma instância da classe "*Academic_Research_Institution*". Com isso é possível verificar que este portal é restrito e fornece apenas informações específicas de uma única instituição (IME), o que torna desnecessária a navegação e a procura de instâncias desta classe. Assim, a navegação deverá ser iniciada pela próxima classe, ou seja, "*Institution_Units*". Para esta classe o portal IME não apresenta *links* com alguma similaridade (subprocesso A), i.e, não apresenta *links* prioritários, e por isso, os demais deverão ser visitados. Durante a navegação pelas páginas destes *links* é verificado se estas contêm tabelas, listas e outros *links*, com alguma similaridade às instâncias da classe "*Institution_Units*". A página referente ao primeiro *link* (não prioritário) visitado é ilustrada na Figura 5 (a). Nesta figura, o primeiro item com maior prioridade encontrado (Ensino de Pós-Graduação – SD/1) é comparado às instâncias da classe. Ao constatar tal similaridade, este deverá ser extraído e armazenado (subprocesso B). Esse processo é realizado para todos os itens (*links*, listas e tabelas) da página visitada.

Figura 4. Página Inicial para a Navegação

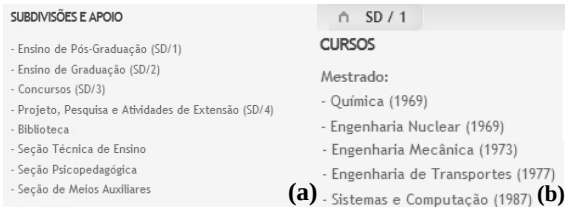


Figura 5. Páginas com Instâncias

Como a navegação pelas páginas do portal é feita de acordo com a OBEDU, após a extração das instâncias da classe "*Institution_Units*", a próxima classe a ter suas instâncias pesquisadas é a "*Education_Program"* e todas as suas subclasses. E por fim, a próxima classe é "*Research_Program"* (Figura 3), repetindo o processo para todas as classes da ontologia até que se esgotem todas as possibilidades de instâncias e o portal tenha sido totalmente visitado. Durante a navegação, a análise da hierarquia (subprocesso C) e a recuperação de pares de instâncias (subprocesso D) também deverão ser efetuadas. São analisadas primeiramente as instâncias das subclasses em relação às instâncias das suas superclasses para eliminar as instâncias repetidas. Como exemplo, podemos citar o programa de Sistemas e Computação, que por ser uma instância da Classe "*Graduation_Program"*, também é uma instância da classe "*Education_Program"*. Assim, a instância da classe mais específica é armazenada e a instância

[4] http://www.comp.ime.eb.br/~posedu
[5] http://www.ime.eb.br/

da classe mais abrangente é descartada. Na recuperação de pares de instâncias, são verificados os relacionamentos entre as instâncias de acordo com as propriedades do objeto. Como observado na Figura 5(b), pode-se dizer que Sistemas e Computação (instância da classe "*Education_Program")* é "*provided_By_Program*", SD/1, que é uma instância da classe "*Institution_Units*". Após esse processo, o usuário fará uma validação (processo iv) dessas informações obtidas e as informações válidas são transformadas em triplas RDF (processo v).

5. Conclusão

Este trabalho apresentou uma arquitetura genérica para a população de uma ontologia de domínio necessária para a atualização automática de um portal semântico. Dentre os módulos constituintes dessa arquitetura, foi dada ênfase ao módulo de Extração de Informações, componente fundamental no processo de atualização de portais. Um estudo de caso no domínio de educação permitiu ilustrar como conteúdos de um portal acadêmico na web aberta podem ser recuperados e integrados a uma ontologia básica de domínio, de modo a prover subsídios para posteriormente popular um portal semântico. Como etapa adicional desse trabalho, pretendemos especificar algumas métricas para validar os resultados obtidos, bem como realizar uma avaliação de usabilidade da ferramenta para identificar possíveis melhorias.

6. Referências

Berners-Lee, T. I.; Hendler, J.; Lassila, O. R. (2001). "The Semantic Web". Scientific American Magazine.

Castaño, A. C. (2008) "Populando ontologias através de informações em HTML - O caso do Currículo Lattes", Dissertação de Mestrado. Universidade de São Paulo. São Paulo, SP.

Kurki, J.; Hyvönen, E. (2010) "Collaborative Metadata Editor Integrated with Ontology Services and Faceted Portals.", Heraklion, Grécia.

Lachtim, F.A.; Cavalcanti, M.C.; Moura, A.M. (2009) "Ontology Matching for Dynamic Publication into Semantic Portals", Journal of the Brazilian Computer Society, ISSN: 0104-6500, vol 15. págs 27- 43, Março.

Lachtim, F.A; Ferreira, G.; Gama, R.; Moura, A.M.; (2009) Cavalcanti, M.C. "POSEDU: a Semantic Educational Portal", IEEE Multidisciplinary Engineering Education Magazine, Vol. 4, Nº 3.

Lachtim, F.A. (2008) "Organização e Instanciação Automática de Conteúdos em Portais Semânticos", Dissertação de Mestrado. Instituto Militar de Engenharia. Rio de Janeiro, RJ.

Hyvönen, E.; Mäkelä, E.; Kauppinen, T; Alm, O.; Kurki, J.; Ruotsalo, T.; Seppälä, K.; Takala, J.; Puputti, K.; Kuittinen, H.; Viljanen, K.; Tuominen, J.; Palonen, T.; Frosterus, M.; Sinkkilä, R.; Paakkarinen, P.; Laitio, J.; Nyberg, K. (2009) "CultureSampo - Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user." Indianapolis, USA.

Poka (2011) "A framework for automatic annotation", http://www.seco.tkk.fi/tools/poka/, Abril.

RDF (2011) "Resource Description Framework (RDF): Concepts and Abstract Syntax", http://www.w3.org/TR/rdf-concepts/, Abril.

Suominem, O.; Hyvönen, E.; Viljanen, K.; Hukka, E. (2009) "HealthFinland - a National Semantic Publishing Network and Portal for Health Information". Finlândia.

Vera (2011) "Validation and quality assistant for Semantic Web data", http://www.seco.tkk.fi/services/vera/, Abril.

Viljanen, K.; Tuominen, J.; Hyvönen, E. (2010) "A Network of Ontology Repositories". Submitted for review.

A relação de meronímia no domínio jurídico: um estudo visando sua inserção em uma ontologia jurídica

Thaís D. Minghelli

Rua João Neves da Fontoura, 716/203. Centro. São Leopoldo/RS - Brasil.

minghelli.adv@terra.com.br

Resumo. *O presente artigo aborda a relação de meronímia (RM) ou parte-todo para fins de representação do domínio jurídico, mais especificamente do Direito Processual Penal (DPP). Tendo em vista a grande demanda por sistemas cada vez melhores no que tange à recuperação da informação, o grande número de sites jurídicos, bem como a relevância social em aclarar a linguagem jurídica para o leigo, ontologias têm se revelado úteis neste sentido. Por isso, objetiva-se aqui refletir de que forma os conceitos do DPP podem ser explicitados sob a luz da relação parte-todo, visando inserir os resultados da pesquisa em uma ontologia.*

1. Introdução

Este artigo apresenta o andamento da pesquisa de mestrado, a qual aborda a RM no domínio jurídico, mais especificamente, no DPP para fins de uma futura inclusão em uma ontologia jurídica linguística. A escolha desta relação semântica para a representação do DPP deve-se ao fato deste domínio ser constituído por um conhecimento segmentado em etapas, em que a RM se mostra recorrente, como também de que referida relação em uma ontologia linguística possibilita a associação pela máquina, por exemplo, das partes/fases do DPP a um todo/atividade tal como um jurista é capaz de fazer.

Neste sentido, a máquina é capaz de identificar que as ações, correspondentes a unidades e não a tipos, frisa-se, <depoimento do ofendido>, <inquirição de testemunhas da acusação e da defesa>, <depoimento dos peritos>, <acareação dos envolvidos> etc., *compõem* um todo, uma atividade, qual seja: a <audiência de instrução e julgamento>. Ontologias permitem a recuperação da informação com maior eficiência, bem como a unificação de ordenamentos jurídicos e a desburocratização, visto que elas ensejam o mapeamento de conceitos, o raciocínio automático e o processamento da linguagem natural (PLN). Na União Européia ontologias jurídicas vêm sendo construídas, visando a comparação e unificação da legislação, como também a recuperação da informação. A *Core Legal Ontology* (CLO), a LRI-Core e LOIS são exemplos de iniciativas neste sentido.

Tendo em vista referidos movimentos e a demanda cada vez maior de aperfeiçoar sistemas de busca em *sites* jurídicos, objetiva-se construir uma ontologia jurídica linguística. Para tanto, apresentam-se os seguintes objetivos: (i) apresentar os *corpora*; (ii) discorrer sobre considerações semânticas acerca da RM, trazendo exemplos jurídicos; (iii) definir e distinguir ontologias, léxicos computacionais e ontologias linguísticas e (iv) mencionar a metodologia e a análise dos tipos de merônimos mais pertinentes para o DPP, objetivando-se verificar em quais dos *corpora*

podem ser encontrados exemplos mais produtivos por meio da utilização de marcadores lingüísticos (ML) no uso da ferramenta *corpógrafo*.

2. Noções prévias

2.1. Os corpora

Os *corpora* correspondem: (i) Ao código de Processo Penal (CPP), ou seja, à norma jurídica aplicada na existência de um processo criminal, em que estão registradas todas as etapas que instruem o desenrolar de uma ação penal, bem como os possíveis participantes do contexto penal e processual penal, as instituições e os documentos. O CPP é composto por artigos 811 artigos, organizados em títulos e capítulos, podendo conter incisos, alíneas e letras. (ii) A dez acórdãos do Tribunal de Justiça do Rio Grande do Sul, coletados por meio de pesquisa orgânica sob as palavras-chaves: <homicídio qualificado apelação>.

2.2. Relação de meronímia ou parte-todo: considerações semânticas

O léxico pode ser estudado sob diferentes perspectivas. Uma delas é a análise por meio de suas relações de sentido com outras palavras, estruturando e regularizando o léxico de uma língua. Dentre os diferentes tipos de relações lexicais, há as ordenadoras do léxico em superordenados/subordinados, em parte-todo e de forma associativa (Saeed, 1997, p.63). Aqui, serão abordadas as relações parte-todo, as quais a priori são noções ontológicas pertencentes à disciplina de Mereologia.

Tendo em vista a interface Linguística do tema, traz-se autores desta linha como Cruse (2000), o qual entende que a meronímia corresponde a uma relação ou conexão entre duas entidades de uma mesma natureza ontológica, aludindo a ideia de inclusão entre dois elementos mutuamente implicados como em X está implicado no sentido de Y, podendo estar dispostos hierarquicamente. Saeed (1997, p. 70) refere que meronímia é um termo usado para descrever relações parte-todo entre dois termos lexicais, em que as partes ou merônimos são unidades de *holônimos*. Conhecida como uma *relação partitiva* é representada por ML, tais como: *parte de, contém, tem, possui*, sendo mais encontrada entre os nominais.

Estudando a relação parte-todo entre pares de unidades lexicais, percebem-se características peculiares a esta relação semântica. Nota-se que algumas partes podem ser mais essenciais diante do todo, tal como em *peça acusatória e processo*. Outras, apesar de comuns, não são obrigatórias, sendo tidas como facultativas como em *rol de testemunhas*, já que uma ação penal pode ser oferecida ainda que sem a menção de testemunhas do fato criminoso. Da mesma forma, a RM pode ser transitiva ou intransitiva, isto é, quando não há o mesmo tipo de RM em todas as partes de uma hierarquia, configurando uma assimetria. É transitiva quando há uma correspondência de *acarretamento* entre as categorias de uma hierarquia. No plano jurídico, exemplifica-se a transitividade: *A votação no Tribunal do Júri* (evento jurídico) *contém leitura dos quesitos. A leitura dos quesitos contém uma ordem. A votação tem uma ordem*. Lyons (1977) verifica que nem sempre a RM é transitiva, sendo a intransitividade gerada pelas variedades de relações parte-todo existentes e presentes em um mesmo silogismo meronímico.

Caracterizando a RM, faz-se a revisão teórica com base em Cruse (2000, p. 154-5). A primeira característica é a **necessidade** de algumas partes ao *holônimo* para sua

adequada formação, configurando uma relação canônica. Em oposição à *necessidade*, hás a **opcionalidade** da parte diante do todo (Cruse,1986, p. 162-3), quando a parte é facultativa na relação. O segundo traço da RM é a **integralidade** da parte ao todo**.** Detecção possível de ser feita pela descrição de como a parte é presa ao todo, sendo que quanto mais difícil esta descrição for, mais integrado parte-todo estarão. No domínio jurídico ilustra-se com o objeto físico *processo,* (*autos* como é chamado tecnicamente) e seus volumes. A **motivação** é outro traço que pode ser identificado na RM, implicando na *função* da parte e permitindo que o todo funcione conforme o fim proposto e inclusive que a parte seja identificada mais facilmente. A **congruência** ou harmonia entre parte e todo pode originar dois fenômenos: a ***supermeronímia*** e a ***semimeronímia***. O primeiro corresponde às partes que são aplicadas a mais de um *holônimo,* como a <apresentação de documentos> é parte da <resposta do réu>, como também é parte da <peça acusatória>. O segundo trata da sobreposição da parte no todo.

No que tange aos tipos de merônimos, Winston et. al. (1987) trazem seis categorias. (i) **Componente-objeto-integral:** relação típica parte-todo, em que o todo inclui a parte, sendo constituído por vários componentes discretos, com limites e funções definidas em relação ao todo. (ii) **Membro-coleção:** parte como um objeto singular e membro de um todo que denota um conjunto. (iii) **Material-objeto:** parte como elemento constitutivo de um objeto concreto, ou seja, à conexão entre a matéria de que é feita e uma entidade concreta. (iv) **Porção-massa:** parte indefinida diante do todo, o qual configura uma entidade mais concreta. (v) **Lugar-área:** parte é o lugar localizado no todo ou área. (vi) **Ação-atividade:** parte como fase de uma atividade/todo, descrevendo diferentes subatividades que compõem uma maior estruturada e cronologicamente organizada. Antecipa-se que dentre aludidas categorias, as mais produtivas para a representação do DPP são a (i), (ii), (v) e (vi), reestudadas quando da análise deste artigo.

2.3. Ontologias, léxicos computacionais e ontologias lingüísticas

O termo ontologia tem sua origem na Filosofia, como o estudo das categorias que compõem o mundo, visando uma classificação universal. Sob o viés tecnológico, as ontologias se diferem da concepção filosófica, apesar de se relacionarem com ela, de certa maneira, já que ambas trabalham com representações de *mundos*. Ontologias ressurgem no contexto da web semântica, criada por Berners Lee, tendo em vista que o conhecimento organizado e relacionado por meio de relações semânticas facilita a recuperação da informação. Ontologias permitem a indexação de textos com maior precisão, permitindo melhores resultados quando da busca virtual. Elas têm o objetivo prático de oferecer estruturas de conhecimento para os sistemas computacionais, possibilitando a resolução de problemas de conhecimento de mundo relacionados ao PLN e o raciocínio lógico automatizado. Como produto tecnológico não buscam a verdade, mas o que existe em um domínio, almejando o melhor funcionamento de um determinado sistema computacional.

Atualmente, ontologias, como estruturas para organização do conhecimento, têm sido estudadas pela Ciência da informação, Linguística Computacional, Inteligência Artificial (IA) com ênfase em aplicações voltadas para o PLN e Web Semântica, por exemplo, visando estruturar e descrever conceitos. Considerando este fato, percebe-se uma imprecisão terminológica e a aproximação de ontologias, léxicos e

ontologias linguísticas. Hirst (2004, apud Prévot, 2010) afirma que *ontologias e recursos lexicais são aparentemente similares o suficiente para que sejam usados de forma intercambiável e combinada,* apesar de possuírem aspectos que os distingam pontualmente. Por tal razão é pertinente apontar em que medida se aproximam e se afastam.

Entende-se que ontologias devem observar a lógica formal. Para a (IA), uma ontologia é uma especificação formal e explícita de uma conceitualização, sendo que o existente é passível de representação (Gruber, 1993). O autor vincula *conceitualização* à especificação, descrição e representação dos termos em uma linguagem de programação, *formal*. *Compartilhado* relaciona ao conhecimento consensualmente compreendido, não restrito a um indivíduo. Para sistemas de informação, uma ontologia é um artefato de engenharia, constituído por um vocabulário específico usado para descrever certa realidade, organizado hierarquicamente e inter-relacionados por relações de suposição. Tais relações descrevem o significado pretendido, considerando-se a finalidade para a qual a ontologia é construída. A hierarquia básica de uma ontologia corresponde a uma taxonomia, relações *é_um* e outras como: de meronímia, *parte_de*, e associativas.

Outra forma de estruturação do conhecimento distinta da ontológica, mas que em vários aspectos se aproxima desta, podendo, por vezes, serem intercambiáveis ou ocorrerem concomitantemente são léxicos computacionais. Aplicações computacionais são um recurso tecnológico em prol do PLN, cujas primeiras aplicações foram para fins de tradução automática, sendo hoje usadas para a recuperação de informação. Atualmente, léxicos estão se tornando recursos robustos com muitas informações linguísticas. WordNets equivalem a importantes bases de dados para o PLN. Por estas razões, a fronteira entre léxicos e ontologias torna-se, cada vez mais tênue. Uma grande diferença entre léxicos e ontologias é que os primeiros descrevem o significado de expressões de linguagem natural; enquanto os segundos descrevem as entidades de um domínio e as relações entre as mesmas. Grosso modo, o que se percebe é um processo de *ontologização* dos léxicos, dando margem a mais um termo afim: ontologias linguísticas. Estas se situariam entre as ontologias formais e os léxicos robustos. Elas emergem da semântica e do léxico, representam um determinado conhecimento por meio de relações semânticas, mas sem o compromisso de se valer da lógica formal. Prévot et. al. (2010) explica que a conceitualização é baseada em critérios linguísticos, mais precisamente em informações encontradas em recursos lexicais.

Neste sentido, ontologias linguísticas caracterizam-se por armazenar apenas conceitos lexicalizados, conceitos expressos por uma ou mais palavras. Sob este viés, uma ontologia linguística corresponde aos sentidos de uma dada língua, aos conceitos compartilhados por uma comunidade lingüística (DI FELIPPO, 2008). O objetivo da pesquisa de mestrado é justamente este: representar o DPP por meio da RM, visando inseri-lo em uma ontologia linguística jurídica, a qual estaria comprometida com a representação do léxico e não com uma linguagem formal.

3. Metodologia

Tendo em vista os *corpora* selecionados (cf.2.2) e as considerações acerca da RM (cf.2.3) pressupõe-se que: os tipos de merônimos (i), (ii), (v) e (vi) de Winston et.al (1987) são pertinentes para a representação do DPP. Assim, objetiva-se, verificar em

quais dos *corpora* podem ser encontrados exemplos mais produtivos por meio da utilização de ML como: tem, composto, constituído, fase, peça, contém, bem como frases preposicionais com preposições como "de" ("do", "da", por exemplo) e "em" ("no", "na"), fazendo uso de concordanciador. Vale esclarecer que neste artigo trouxeram-se apenas os ML que ensejaram resultados, considerando ainda a concisão do artigo. Estabelece-se, assim, a metodologia para a realização da análise: (i) pautado nos pressupostos estudados, testaram-se os ML que representam relação de parte-todo; (ii) utilizando o *corpógrafo*, testaram-se os ML nos *corpora*; (iii) analisaram-se qual *corpus* trouxe exemplos que melhor representassem o domínio do DPP e (iv) apresentaram-se, neste artigo, somente os pares de parte-todo que trouxeram resultados de meronímia para inclusão em uma ontologia jurídica lingüística.

4. Análise dos dados

Testaram-se os ML acima mencionados; todavia poucos se mostraram produtivos em ambos os *corpora*. Em decorrência do limite de páginas, mencionaremos somente um exemplo de cada retorno pertinente para a representação da RM no DPP, procurando um enquadramento nas tipologias de Winston et. al. (1987) retro-referidas:

ML *composto:* recuperou 01 exemplo apenas no *corpus* do CPP e nenhum no conjunto de 10 acórdãos. Ex. O Tribunal do Júri é **composto** por 1 (um) juiz togado, seu presidente e por 25 (vinte e cinco) jurados que serão sorteados dentre os alistados, 7 (sete) dos quais constituirão o Conselho de Sentença em cada sessão de julgamento.

ML *fase:* recuperou 11 concordâncias do CPP e 3 dos acórdãos, sendo que todos expressam o mesmo tipo de RM. Os retornos obtidos ilustram a categoria ação-atividade, pois aludem às subatividades de um evento maior. Ex. `O exame poderá ser ordenado na do inquérito, mediante representação da autoridade policial ao juiz competente.`

ML *conterá:* recuperou 11 concordâncias do CPP e 0 dos acórdãos, sendo todas as 11 tipologias de componente-objeto-integral, porquanto elementos constitutivos de um todo em sua forma mais típica. Ex. `A representação conterá todas as informações que possam servir à apuração do fato e da autoria.`

5. Considerações finais

Tendo em vista o limite de páginas imposto, resumem-se as considerações finais neste parágrafo. O presente artigo revisou a literatura quanto à RM, situou ontologias, léxicos computacionais e ontologias linguísticas. Na análise, comparou dois *corpora* do domínio especializado, almejando testar se a busca automática por merônimos seria produtiva. A análise mostrou que: (i) provavelmente pela especialidade do domínio jurídico, a busca automática não restou produtiva, apesar de ter apresentado alguns resultados, conduzindo à conclusão de que a busca por merônimos deverá ser manual no caso dos corpora jurídicos; (ii) o *corpus* do CPP trouxe mais resultados; (iii) poucos ML trouxeram resultados precisos, a maioria deles esbarrou na polissemia dos termos jurídicos, como o sentido de *peça* como documento jurídico escrito, *parte* como autor ou réu, *tem* ou *possui* no sentido de posse e *constituído* como os poderes outorgados ao defensor para representar em juízo e (iv) a categoria ação-atividade de Winston et. al. (1987) foi o tipo de meronímia mais recorrente no DPP.

Referências Bibliográficas

BERTOLDI, A. A semântica dos adjetivos: como e por que incluí-la em uma ontologia de domínio jurídico. Dissertação. UNISINOS, São Leopoldo, 2007.

BREUKER J, and WiINKEL, R. *Use and reuse of legal ontologies in knowledge engineering and information management*. ICAIL03 Wks on Legal Ontologies and Web-based Information Management, Edinburgh, http://lri.jur.uva.nl/~winkels/legontICAIL2003.html, 2003. **(LRI-Core)**

CRUSE, D. A. *Lexical Semantics*. Cambridge: Cambridge University Press, 1986.

CRUSE, D. A. *Meaning in language: an introduction to semantics and pragmatics*. Oxford: Oxford University Press, 2000.

DI FELIPPO, Ariani. (2008). *Ontologias lingüísticas aplicadas ao processamento automático das línguas naturais: o caso das redes wordnets*. In: Magalhães, J. S.; Travaglia, L. C. (Orgs). Múltiplas perspectivas em Linguística. Uberlândia: Edufu, 2008. ISBN 978-85-7078-200-7.

GANGEMI, Aldo. et al. *Some ontological tools to support legal regulatory compliance, with a case study***.** Laboratory for Applied Ontology, ISTC-CNR, Rome Italy** Università per Stranieri, Perugia, Italy *** ITTIG-CNR, Florence, Italy. s/a.

GANGEMI, A.; SAGRI, M. T. e TISCORNIA, D. *A constructive framework for legal ontologies*. In: BENJAMINS, V. R. et al. (Eds.). Law and the Semantic Web: Legal ontologies, methodologies, legal information retrieval, and applications, LNAI (3369). Berlin/Heidelberg: Springer-Verlag, 2005, p.97-124.

GRISHMAN, R. CALZOLARI, N. "Lexicons". In COLE, R. (Ed.). *Survey of the state of the art in human language technology.* Studies in Natural Language Processing. York: Cambridge University Press. 1997. p 368-370.

GRUBER, Thomas R. *A Translation Approach to Portable Ontology Specifications*. In: Knowledge Acquisition, **5**(2):199-220, 1993.

MARRAFA, Palmira. *WordNet do Português: uma de dados de conhecimento linguístico.* Instituto Camões. 2001.

MILLER, George A. et al. *Introduction to WordNet: An On-line Lexical Database.* 1993.

PRÉVOT, L. et. al. *Ontology and the lexicon: a multidisciplinary perspective.* In: C. Huang. et al. *ontology and the lexicon: a natural language processing perspective.* New York: Cambridge University Press, 2010, p.3-24.

RIEMER, Nick. *Introducing semantics*. Cambridge University Press. 2010.

SAEED, John I. *Semantics*. Blackwell Publishers Inc. 1997.

Sites: Corpógrafo: http://www.linguateca.pt

CLO: www.ittig.cnr.it

LOIS: www.loisproject.org

WordNet: Disponível em: http://wordnet.princeton.edu/perl/webwn?s=word-you-want. Acessado em: 26/10/2010.

EuroWordNet: Disponível em: http://www.illc.uva.nl/EuroWordNet/. Acessado em 27/10/2010.

SOWA, J. F. (2006). Building, sharing and merging ontologies. Disponível: http://www.jfsowa.com/ontology/ontoshar.htm. acesso em 29/10/2010.

Ontologias sobre segurança da informação em biomedicina: tecnologia, processos e pessoas

Luciana Emirena dos Santos Carneiro, Maurício Barcellos Almeida

Escola da Ciência da Informação – Universidade Federal de Minas Gerais – UFMG
Avenida Antônio Carlos, 6627– Campus Pampulha - 31.270-901–Belo Horizonte – MG

lucianaemirena@gmail.com, mba@eci.ufmg.br

Abstract. *An issue concerning the organizational information security is the lack of standards to describe incidents. One of the first initiatives to face attacks and failures may be the creation of a uniform vocabulary. Ontologies are an alternative to organize such vocabulary. This paper advocates that information security guidelines are effective insofar as they are able to encompass three perspectives – namely, technology, people, and processes – and that such perspectives are present within ontologies development process. We describe the preliminary terminological stage of an information security ontology, part of an ongoing research, as well as its future uses.*

Resumo. *Uma questão sobre segurança da informação nas organizações é a falta de padronização para descrever incidentes. Um vocabulário uniforme é o primeiro passo responder as tentativas de ataque e falhas. Ontologias são uma alternativa para organizar tal vocabulário. Esse artigo advoga que políticas de segurança da informação são efetivas na medida em que abrangem três perspectivas – tecnologia, pessoas, e processos – e que tais perspectivas estão presentes no processo de desenvolver ontologias. Descreve-se estágio terminológico preliminar de ontologia para segurança da informação, parte de pesquisa em andamento, bem como suas possibilidades de uso.*

1. Introdução

A evolução dos sistemas de informação tem possibilitado às empresas ganhos em mobilidade e conectividade. A descentralização promovida pelas redes, entretanto, tem exigido mais atenção à gestão e ao controle, em um conjunto de esforços que se convencionou chamar de "segurança da informação".

Os investimentos em segurança da informação tem sido crescentes, mas existem dificuldades em definir: o que deve ser protegido, qual o nível de proteção necessário, e quais as ferramentas utilizar no ambiente corporativo. A TI tem meios para solucionar parte do problema de segurança, mas o elemento humano ainda representa grande parte das ocorrências e falhas de segurança [Colwill, 2010].

O presente artigo apresenta pesquisa em andamento na qual se busca os requisitos para definir segurança da informação na perspectiva de tecnologia, pessoas, e processos. Usa-se uma abordagem baseada em ontologias para classificar informações sobre segurança, cujos resultados correspondem ao estágio terminológico preliminar de uma ontologia de domínio sobre segurança da informação na área de saúde, a qual ainda requer decisões de natureza ontológica, criação de restrições, validação, implementação, dentre outros. Tais ações estão planejadas para a sequencia da pesquisa.

O restante do artigo está dividido conforme segue. A seção 2 apresenta uma visão geral da segurança da informação e a seção 3 detalha a segurança do ponto de vista das três perspectivas mencionadas. A seção 4 descreve a metodologia de pesquisa, a seção 5 apresenta resultados parciais e a seção 6 apresenta considerações finais.

2. Segurança da informação: uma visão geral

Segurança da informação é uma questão multifacetada, composta por diversas variáveis interagindo em um único ambiente. Uma empresa demanda confidencialidade da informação sem, entretanto, perder a disponibilidade frente aos riscos, ameaças e vulnerabilidades. A segurança da informação abrange a totalidade dos elementos de negócio corporativos: a busca de melhorias em processos que garantem a qualidade e competitividade; o aprendizado e a produção de conhecimento organizacional; a criação de modelos e o uso das informações; detecção e prevenção, assim como documentação de potenciais riscos, ameaças e vulnerabilidades.

A TI é quesito fundamental em qualquer solução para a segurança da informação. Os sistemas, contudo, são projetados, implementados e operados por pessoas. São pessoas que proporcionam segurança física, concedem acesso aos sistemas, causam, relatam e gerenciam a resposta das empresas frente às violações e incidentes de segurança [Lacey, 2009].

Segundo Sveen, Torres e Sarriegi (2009), o desenvolvimento de um plano estratégico de segurança da informação efetivo depende dos aspectos: i) pessoas, como formadoras da cultura organizacional; ii) processos, como condutores do fluxo informacional; e iii) tecnologias, ferramentas que sustentam processos e necessidades dos usuários.

3. Fatores intervenientes em segurança da informação

A concepção sobre o que é segurança da informação tem evoluído e não mais se restringe apenas à questão técnica. A segurança da informação está atrelada ao negócio da empresa através da necessidade de proteção dos ativos informacionais. Esses ativos informacionais não incluem apenas informações valiosas em repositórios da empresa, ou em suas marcas, mas também o valor fornecido pelo know-how, expertise, habilidades e relacionamentos incorporados na rede corporativa, dentro e fora de seus limites físicos. Os ativos informacionais envolvem assim pessoas, executando processos, em geral, como uso de tecnologia.

3.1. Segurança e pessoas

Incluir pessoas nos estudos de segurança significa dar atenção devida à subjetividade inerente ao seres humanos, suas relações e seu comportamento informacional nas organizações que tanto influencia a gestão e a tecnologia.

O elemento “pessoas” gera vulnerabilidade, uma vez que a falta de conhecimento ou treinamento resulta em condutas inapropriadas às ações de segurança. Colaboradores das organizações, intencionalmente ou por negligência, são a maior ameaça à segurança da informação [Van Niekerk e Solms, 2010]. A segurança depende tanto do conhecimento humano quanto de sua cooperação. A falta de conhecimento é tratada através da educação ou treinamento, enquanto a falta de cooperação é abordada através da promoção de uma cultura de que incentive preocupação com a segurança.

As organizações não podem proteger a integridade, confidencialidade e disponibilidade das informações em ambiente de sistemas em rede, sem garantir que cada pessoa envolvida compreenda suas responsabilidades, e seja treinada para realizá-las.

3.2. Segurança e processos

A adoção de políticas, procedimentos, normas e diretrizes relativas à segurança em organizações, busca tornar claro o comportamento esperado e as regras a seguir. Nesse sentido, uma série de fatores contribuem para segurança ineficaz: falta de especificação e de documentação, incapacidade interna de criar políticas efetivas, e a falta de mecanismos de execução [Kraemer, Carayon e Clem, 2009]. Incidentes decorrentes da ineficácia dos controles são minimizados através de rotinas e instruções acessíveis.

Um fator-chave no sucesso de uma política de segurança é a implantação dos controles de segurança, incluindo o acompanhamento, sanções e recompensas de colaboradores, proporcionais aos potenciais riscos envolvidos [Sianes, 2006].

3.3. Segurança e tecnologia

A necessidade de proteção dos ativos informacionais das empresas faz com que a tecnologia ganhe uma posição de destaque. Existem diversos tipos de controles para auxiliar na gestão da segurança, na limitação dos incidentes e na violação de segurança [Sveen, Torres e Sariegi, 2009]: check-lists, análise de risco, avaliação e métodos de detecção, dispositivos biométricos e de bloqueio, antivírus, firewalls, criptografia, permissões na rede, auditorias, dentre outros.

A ISO/IEC-15408-1 (2005) é a principal referência para avaliação de atributos de segurança em produtos de TI. Esta norma estabelece um critério comum para a avaliação, possibilitando que o resultado seja significativo para audiências variadas.

No âmbito da Internet, cabe destacar o papel do Computer Emergency Response Team / Coordination Center (CERT/CC), cujo objetivo é centralizar a coordenação de respostas à incidentes de segurança.

4. Metodologia de pesquisa

Diversas iniciativas de uso de ontologias em segurança da informação estão disponíveis na literatura [Raskin et al, 2001] [Martiniano e Moreira, 2007] [Fenz et al, 2007] [Ekelhart et al, 2006]. Apresenta-se a seguir um conjunto de procedimentos utilizado em uma organização de saúde para criar a ontologia relativa à segurança. O processo foi divido em três etapas: i) organização da informação registrada; ii) organização da informação especializada; iii) terminologia para ontologia.

Etapa (1) – organização da informação registrada em documentos e em sistemas

- Organização: classificam-se os documentos a partir de seu conteúdo e de sua proveniência, registra-se sua tipologia e seu ciclo de vida, elegem-se os vitais;
- Padronização: acrescenta-se uma folha de rosto a cada documento, na qual são registrados dados como autor, data de emissão, etc;
- Treinamento: os colaboradores são orientados sobre como classificar documentos assim que este é produzido;

- Relatórios: são gerados a partir de registros e dados provenientes de sistemas utilizados no setor, os quais são tratados da mesma forma que os documentos.

Etapa (2) – organização da informação especializada fornecida por pessoas

- Aquisição de conhecimento: obtém-se com os colaboradores e especialistas as informações sobre suas atividades, documentos que utilizam, conceitos e relações relevantes para o entendimento de suas práticas;
- Técnicas: são utilizadas técnicas bem consolidadas, como entrevistas, grade repertórios, ordenamento de cartões, dentre outras;
- Relatórios: o material obtido, sejam de cunho administrativo ou científico, é registrado em sínteses de entrevistas e planilhas;

Etapa (3) – elaboração da terminologia

- Organização preliminar: os dados obtidos são organizados em uma lista composta por termos obtidos nas etapas 1 e 2;
- Estágio terminológico: a lista é organizada agrupando-se substantivos candidatos a conceitos, e verbos candidatos a relações na ontologia; cabe destacar que tais correspondências – verbo ⇔ relação, e substantivo ⇔ classe – são mais simples do que aquelas requeridas para construir a ontologia, mas atendem a demanda por organização preliminar do vocabulário nessa etapa da pesquisa.

Os processos organizacionais e as práticas dos colaboradores registradas em documentos e em sistemas, os quais correspondem à grande parte da informação da organização, são classificados e relacionados entre si. O estágio terminológico, produto parcial da pesquisa em andamento, é parte da atividade de desenvolvimento de ontologia. A atividade de desenvolvimento de ontologias abrange etapas de aquisição e modelagem correspondem à formas de incrementar a comunicação humana [Almeida e Barbosa, 2009]. A ontologia será, após sua consecução, a referência única para classificar dados em uso na organização relativos a segurança da informação.

5. Resultados parciais

Os termos obtidos pela metodologia descrita na seção 4 caracterizam a organização preliminar de dados e resultam no que se denominou aqui de "estágio terminológico". A Tabela 1 resume as entidades consideradas básicas e traz uma descrição de seu significado obtida no âmbito da organização de saúde.

Tabela 1 – Exemplos de termos e definições da ontologia de segurança

Entidade	Descrição
Organização	Organização é uma entidade social composta por recursos materiais e humanos, e caracterizada por objetivos, procedimentos de controle e limites. Ex. pública, privada.
Atributo de segurança	Atributo de segurança caracteriza um ativo e diz respeito a requisitos de segurança sobre tal ativo. Pode ser um atributo de confidencialidade, ou de integridade.
Ativo	Ativo é um bem da organização, seja físico ou imaterial, utilizado

	pelos seus membros para alcançar os objetivos estipulados. Ex. um, um sistema, um documento.
Controle	Controle é um procedimento sistematizado para atenuar vulnerabilidades, bem como para estabelecer medidas preventivas e corretivas com vistas à proteção de ativos.
Ameaça	Ameaça é uma possibilidade de dano aos ativos da organização, que afeta atributos de segurança e explora vulnerabilidades. Ex. de origem humana ou natural.
Vulnerabilidade	Vulnerabilidade é a situação caracterizada por falta de medidas de proteção e que possui um grau de severidade associado. Pode ser administrativa, técnica ou física.

A Figura 1 apresenta a organização de termos correspondente ao estágio terminológico preliminar, resultado parcial de pesquisa. Enfatiza-se que o estágio de pesquisa aqui apresentado requer ainda considerações para que se possa obter a ontologia. A organização de termos e definições se restringe, nessa etapa, a uma visão terminológica que ainda carece de considerações adicionais de natureza ontológica. O conjunto terminológico é composto por 165 termos representativos de conceitos no domínio da segurança da informação, bem como classes. Os termos estão distribuídos da seguinte forma: 13 tipos de ativo (por ex., móvel, sistema), 3 tipos de organização (por exemplo, pública), 11 tipos de atributos de segurança (por ex. confiabilidade), 51 tipos de ameaça (por ex., controle de operações em servidores), e 40 tipos de vulnerabilidades (por ex. término de contrato de trabalho de colaborador).

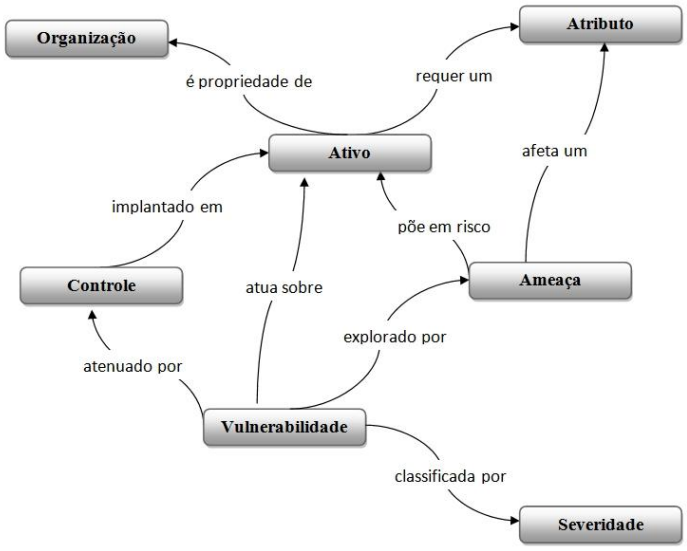


Figura 1: Organização preliminar de termos básicos em forma de rede semântica

6. Considerações finais

O presente artigo aborda a necessidade de se tratar a Segurança da Informação sob a ótica da tríade "pessoas – processos – tecnologias" no contexto corporativo. Através das

ontologias pretende-se registrar, classificar e relacionar as informações de segurança informacional através de uma abordagem sócio-técnica em instituições de saúde.

Entende-se que através do mapeamento das práticas de segurança da informação dos colaboradores da área biomédica, seu registro e documentação, e conseqüente classificação e relacionamento das entidades apuradas, promover-se-á um mecanismo de criação de conhecimento organizacional e desenvolvimento para processos e sistemas.

O desenvolvimento de ontologias é uma oportunidade para unir perspectivas e integrar pessoas, processos e tecnologias de forma equitativa. Fomenta a criação, aquisição e compartilhamento de conhecimento, bem como a aprendizagem organizacional. Os resultados esperados apontam para um produto de informação, uma base de conhecimento, que auxiliará o desenvolvimento de sistemas e políticas no contexto empresarial em que for aplicada.

Na continuidade da pesquisa, pretende-se reavaliar ontologicamente os dados organizados no estágio preliminar de termos, buscando alinhamento com ontologias de alto nível e, em última instância, a redução da ambigüidade.

7. Referências

Almeida, M. B.; Barbosa, R. R. (2009). Ontologies in knowledge management support - a case study. In: Journal of American Society of Information Science and Technology. vol. 60, n.10, p. 2032-2047.

Colwill, C.(2010). Human factors in information security: The insider threat e Who can you trust these days? In: Information Security Technical Report, p. 01-11.

Ekelhart, A. et al.(2006) Security ontology; simulating threats to corporate assets, http://www.springerlink.com/index/w530v5081301j833.pdf, July.

Fenz, S. et al.(2007) Information security fortification by ontological mapping of the ISO/IEC 27001 Standard, http://www.ifs.tuwien.ac.at/node/4274, November.

Lacey, D.(2009). Managing the human factor in information security. Wiley.

Kraemer A. S.; Carayon, P.; Clem, J. (2009) Human and organizational factors in computer and information security: Pathways to Vulnerabilities. In: Computer e Security, v.28, p. 509-520.

Martiniano, L. A. F.; Moreira, E. S. (2007) An OWL-based security incident ontology, http://protege.stanford.edu/conference/2005/submissions/posters/poster-martimiano.pdf, November.

Raskin, V. et al. (2001) Ontology in information security; a useful theoretical foudation and methodological tool, http://portal.acm.org/ft_gateway.cfm?id=508180, August.

Sveen, F. O.; Torres, J. M.; Sarriegi, J. M. (2009). Blind Information Security Strategy. In: International Journal of Critical infrastructure Protection, v.2, p.95-109

Sianes, M. (2005). Compartilhar ou proteger conhecimentos? In: Gestão Estratégica da Informação e Inteligência Competitiva. São Paulo.

Van Niekerk, J. F.; Von Solms, R. (2010). Information Security Culture: A management perspective. In: Computer & Security, n.4,v.29, p.476.

Uma Ontologia para Gestão de Segurança da Informação

Paulo Fernando da Silva, Henrique Otte, José Leomar Todesco, Fernando A. O. Gauthier

Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (EGC) – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

{paulofernando@furb.br, otte@stela.org.br, tite@inf.ufsc.br, gauthier@egc.ufsc.br

Abstract. *This article presents some problems of knowledge management in management information security organizations and suggests ontologies as part of solving these problems. The article describes the construction of a specific ontology for information security using the methodology NeOn and discusses the use of this ontology in the management information security environment.*

Resumo. *Este artigo apresenta alguns problemas de gestão do conhecimento em organizações de consultoria de gestão de segurança da informação e sugere ontologias como parte da solução destes problemas. O artigo descreve a construção de uma ontologia para gestão de segurança da informação utilizando a metodologia NeOn e discute o uso desta ontologia no cenário de gestão de segurança da informação.*

1. Introdução

A gestão de segurança da informação é realizada por consultores internos ou externos em uma organização com o objetivo de identificar o grau de segurança de um ambiente corporativo e propor controles tecnológicos ou administrativos para a redução dos riscos de incidentes de segurança da informação neste ambiente.

Geralmente o conhecimento necessário para a realização de um projeto de gestão de segurança da informação está descrito em normas técnicas, como a ISO 27001 (conhecimento explícito), ou internalizado na mente dos consultores de segurança da informação (conhecimento tático), sendo que neste segundo caso a qualidade do projeto de gestão de segurança da informação depende da experiência e prática do consultor. Este fator torna-se um problema na medida em que empresas de consultoria em segurança da informação não conseguem manter a mesma qualidade de atendimento em todo o seu corpo de consultores, ou seja, um consultor com mais experiência ou conhecimentos específicos poderá desempenhar um trabalho diferenciado frente a outros consultores [Kim, 2007].

Considerando também o problema da dispersão de conhecimento (em diversas normas técnicas, políticas, boas práticas e na mente dos consultores) nas organizações de consultoria em gestão de segurança da informação, a construção de uma especificação explícita e formal para o gerenciamento deste conhecimento seria um grande avanço rumo à gestão do conhecimento nestas organizações.

As ontologias visam à definição de semântica para representação do conhecimento em um dado contexto [Bachimont, 2002]. Uma ontologia aplicada à gestão de segurança da informação poderia contribuir para a gestão do conhecimento em empresas de consultoria de gestão de segurança da informação, servindo de auxílio para os processos de aquisição, representação, armazenamento e compartilhamento de conhecimento obtido a partir de normas e consultores de segurança da informação.

O objetivo inicial deste artigo é mostrar a construção de uma ontologia para gestão de segurança da informação seguindo a metodologia NeOn de desenvolvimento de ontologias. Futuramente espera-se que esta ontologia possa ser utilizada na gestão de conhecimento em organizações de consultoria de segurança da informação.

A seção 2 deste artigo apresenta os conceitos de segurança da informação que serviram de base para a construção da ontologia. A seção 3 descreva a construção da ontologia a partir dos conceitos de segurança da informação. A seção 4 descreve resultados obtidos até o momento com o desenvolvimento da ontologia, e a seção 5 apresenta as conclusões e extensões do projeto.

2. A Gestão de Segurança da Informação

Vários elementos compõem a Gestão de Segurança da Informação. Estes elementos permitem a realização de análises de risco, definição de controles de segurança da informação e a melhoria contínua do ambiente. Um elemento essencial na gestão de segurança da informação é o ativo de informação. Ativo de informação é qualquer informação que possua valor para a organização, bem como qualquer outro elemento de infraestrutura que forneça suporte a esta informação, como: hardwares, softwares e ambientes físicos [Campos, 2007].

Para a realização de uma análise de risco de segurança da informação existe a necessidade da definição de Ameaças e Vulnerabilidades relacionadas ao ambiente a ser avaliado. Ameaça é o agente ou condição que realiza um incidente de segurança da informação [ABNT, 2006]. Grupos comuns de ameaças podem ser: invasores internos, invasores externos, eventos naturais e programas maliciosos (*malwares*).

Vulnerabilidades são falhas em potencial existentes nos ativos de informação. As vulnerabilidades podem ser agrupadas de diversas maneiras. Uma forma simples de agrupar ou categorizar as vulnerabilidades ocorre dividindo-as em: deficiências físicas ou deficiências lógicas [ABNT, 2007].

Os conceitos discutidos acima serão modelados em uma ontologia para gestão de segurança da informação (seção 3 deste artigo) com o objetivo de se estabelecer uma conceituação clara e explícita dos mesmos.

3. Ontologia para Segurança da Informação Utilizando Metodologia NeOn

A partir de conceitos gerais de gestão de segurança da informação, foram modeladas classes de indivíduos para uma ontologia de segurança da informação. Estas classes foram modeladas utilizando-se a ferramenta NeOn e a metodologia NeOn de construção de ontologias [Suárez-Figueroa, 2008]. A Figura 1 apresenta as classes resultantes a partir dos conceitos de segurança da informação levantados. Pode-se observar a criação

das classes Ameaça, Vulnerabilidade e AtivoInformação, bem como suas respectivas sub-classes.

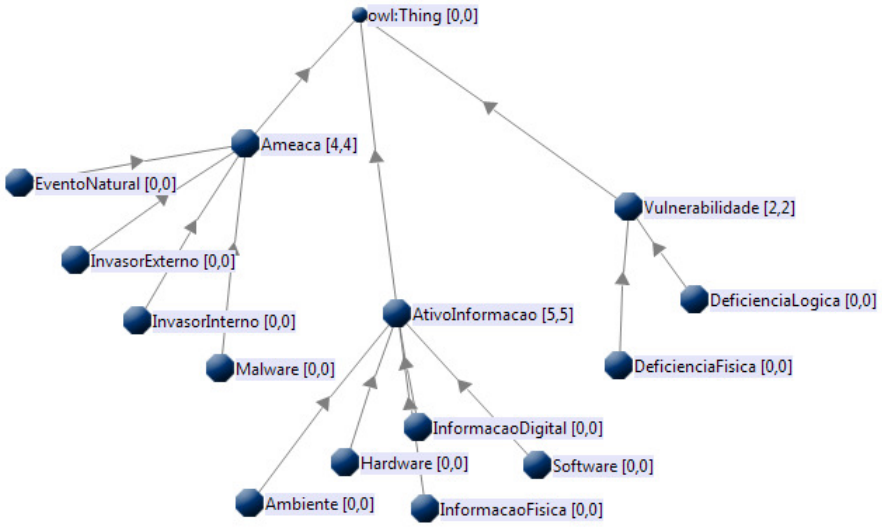


Figura 1. Modelagem das classes da ontologia

Após a definição das classes de indivíduos da ontologia, é necessário estabelecer a relação entre elas. Conforme os conceitos de gestão de segurança da informação, um incidente ocorre quando uma ameaça explora uma vulnerabilidade existente em um ativo de informação. Outro conceito importante é o fato de que ativos de informação podem estar localizados dentro de outros ativos de informação, por exemplo, um documento digital importante para a organização está localizado em um servidor, que por sua vez está localizado em um *Data Center* (ambiente físico).

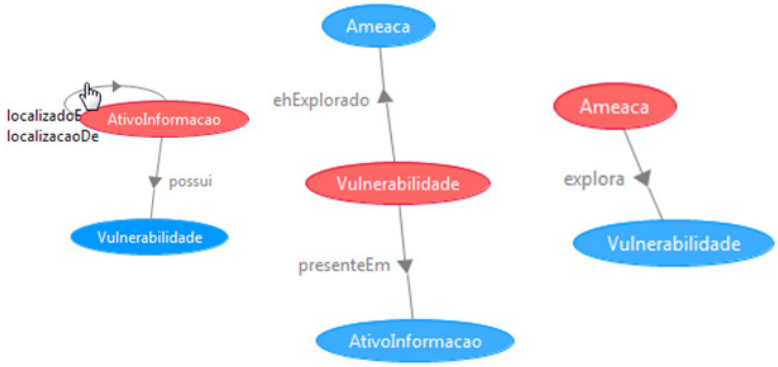


Figura 2. Visualização dos relacionamentos

Com base no exposto acima, foram criados os seguinte relacionamentos na ferramenta NeOn: *localizadoEm, localizaçãoDe, possui, presenteEm, ehExplorado, explora*. A Figura 2 apresenta a relação entre as classes de indivíduos com base nos relacionamentos criados, onde se pode observar que um ativo de informação está localizado em outro ativo de informação ao mesmo tempo em que também é a localização de outro ativo de informação. Uma vulnerabilidade está presente em um

ativo de informação e é explorada por uma ameaça. E uma ameaça possui o relacionamento de explorar um ativo de informação.

Além das definições de relacionamentos, também foram modeladas restrições entre as classes da ontologia. Por exemplo, o relacionamento *localizadoEm* não pode ocorrer entre as classes Hardware e Software, pois não faz sentido um hardware estar localizado em um software. O mesmo ocorre entre as classes Ambiente e Hardware, por exemplo. Existem restrições também entre o relacionamento de Ameaça com Vulnerabilidade, por exemplo, uma vulnerabilidade física não pode ser explorada por uma ameaça da classe *Malware*, bem como uma vulnerabilidade lógica não pode ser explorada por uma ameaça da classe EventoNatural.

Após a definição das classes, dos relacionamentos e das restrições, a ontologia foi populada com vários indivíduos representativos das classes. Conforme apresenta a Figura 3, foram criados indivíduos para todas as subclasses de ativos de informação, ameaças e vulnerabilidades previamente modeladas.

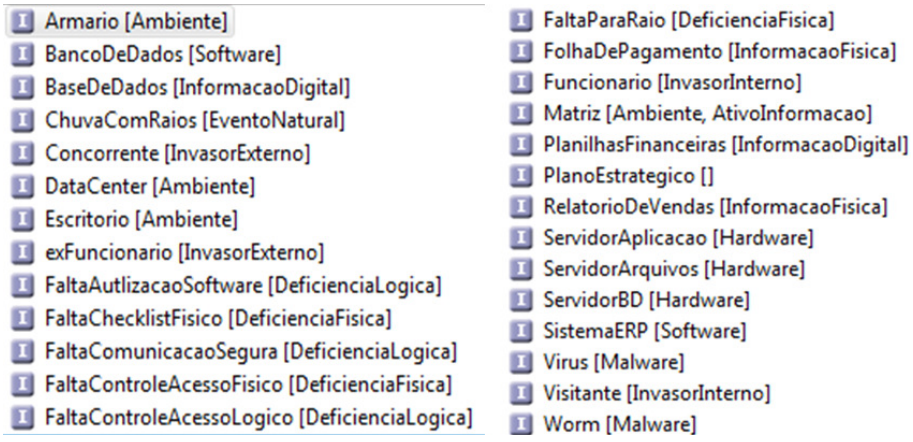


Figura 3. Criação dos indivíduos da ontologia

Uma vez que os indivíduos estão cadastrados, estes já recebem influência dos relacionamentos e das restrições modeladas na ontologia, ou seja, a ontologia já fornece semântica para os indivíduos, a partir dos relacionamentos das classes e das restrições.

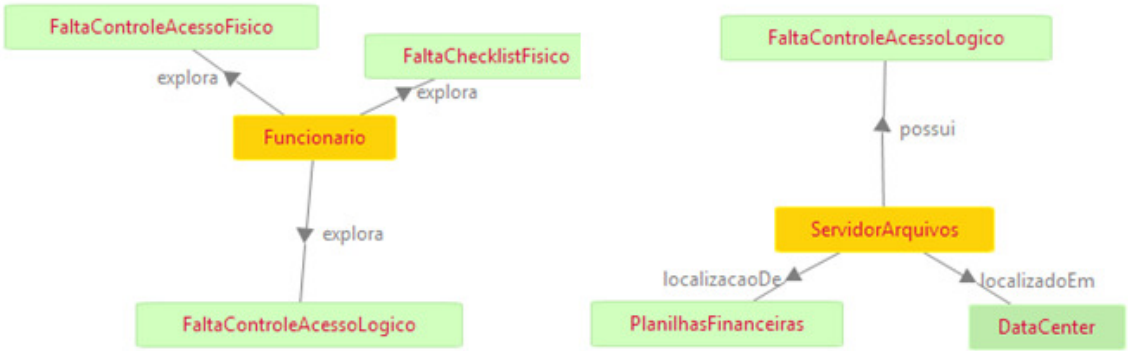


Figura 4. Visualização dos relacionamentos

A Figura 4 (extraída diretamente da ferramenta NeOn) apresenta um exemplo de relacionamento de indivíduos inferido através da ontologia modelada. Através da ontologia e dos indivíduos cadastrados pode-se concluir que o funcionário é uma

ameaça no cenário de segurança da informação que explora as vulnerabilidades de falta de controle de acesso físico, falta de *checklist* físico e falta de controle de acesso lógico. Da mesma forma o servidor de arquivos é um ativo de informação que possui a vulnerabilidade de falta de controle de acesso lógico, também é a localização de outro ativo de informação que é a planilha financeira e está localizado no Data Center da empresa.

Esta seção apresentou a modelagem de uma ontologia para gestão de segurança da informação e a criação de indivíduos representativos para as classes desta ontologia.

4. Resultados e Discussão

Com uma ontologia para gestão de segurança da informação desenvolvida sob a metodologia NeOn de desenvolvimento de ontologias é possível aplicar qualquer ferramenta de inferência compatível com o padrão OWL. Outra vantagem é a possibilidade de integração desta ontologia com outras ontologias de outros domínios, por exemplo: esta ontologia que se fundamenta na ISO 27001 poderia ser integrada com outras ontologias desenvolvidas com base em outras normas da ISO (ex. ISO 9001), formando assim uma ontologia maior e integrada.

Para demonstrar os resultados e possibilidade a partir do desenvolvimento de uma ontologia para segurança da informação, fez-se uso de uma ferramenta compatível com o NeOn para inferência e consulta sobre ontologias no padrão OWL – o SPARQL.

A partir do SPARQL é possível realizar questionamentos complexos sobre a ontologia e seus indivíduos. A Figura 5 demonstra a execução de uma consulta SPARQL sobre a ontologia modelada e os indivíduos criados neste projeto.

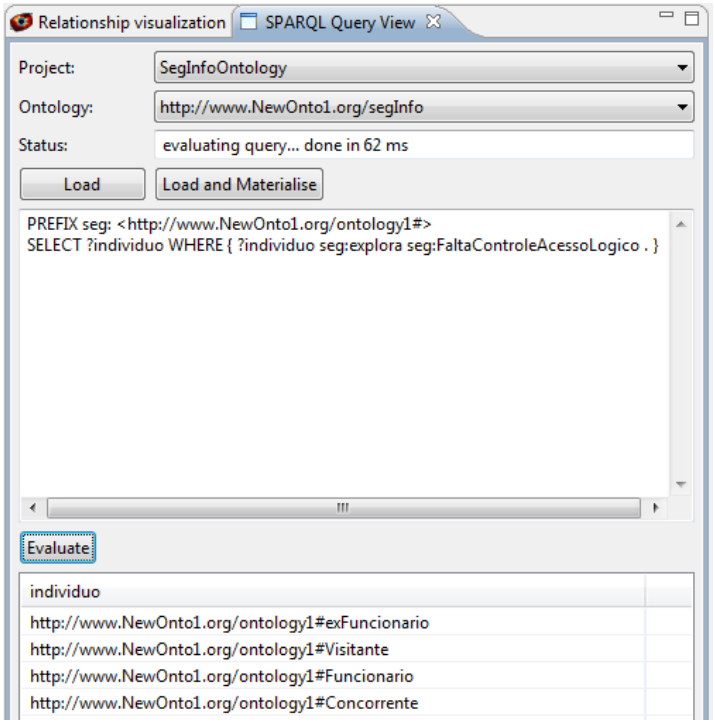


Figura 5. Visualização dos relacionamentos

Foi questionado à ontologia quais são os indivíduos que exploram a falta de controle de acesso lógico, como resultado o SPARQL apresentou os indivíduos: ex-funcionário, visitante, funcionário e concorrente. A partir deste exemplo seria possível realizar qualquer tipo de questionamento com base nas classes e relacionamentos modelados.

O uso do SPARQL neste exemplo foi feito através de um console de consulta, porém no desdobramento deste projeto o SPARQL e outras ferramentas de inferência sobre ontologia poderão ser utilizados em forma de biblioteca dentro de um ambiente mais amplo de suporte à gestão do conhecimento de segurança da informação em empresas de consultoria de gestão de segurança da informação.

5. Conclusão

Este artigo apresentou a concepção, modelagem, população e teste de uma ontologia para gestão de segurança da informação, desenvolvida sob a metodologia NeOn de desenvolvimento de ontologias. Esta ontologia servirá de base para o desenvolvimento de uma ferramenta de suporte à gestão do conhecimento em organizações de consultoria de segurança da informação. A ontologia de gestão de segurança da informação auxiliará na aquisição, representação, armazenamento e compartilhamento de conhecimento relacionado com gestão de segurança da informação.

Como extensão deste trabalho sugere-se: a ampliação da ontologia, com a adição de classes para o suporte de controles de segurança da informação tecnológicos e administrativos; a aplicação desta ontologia em uma arquitetura para gestão do conhecimento em organizações de consultoria de gestão de segurança da informação.

References

ABNT NBR ISO/IEC 27001:2006, Tecnologia da informação - Técnicas de segurança - Sistemas de gestão de segurança da informação – Requisitos.

ABNT NBR ISO/IEC 27002:2007, Tecnologia da informação - Técnicas de segurança - Código de pratica para a gestão de segurança da informação.

Campos, André L. N. Sistemas de Segurança da Informação: controlando os riscos. São Paulo: Visual Books, 2007.

Suárez-Figueroa, M. C.. K. Dellschaft, E. Montiel-Ponsoda, B. Villazón-Terrazas, Z. Yufei, G. Aguado de Cea, A. García, M. Fernández-López, A. Gómez-Pérez, M. Espinoza, M. Sabou. NeOn Deliverable D5.4.1. NeOn Methodology for Building Contextualized Ontology Networks. NeOn Project. http://www.neon-project.org. February 2008.

Kim, Sung-kwan. Silvana Trimi: IT for KM in the management consulting industry. J. Knowledge Management 11(3): 145-155 (2007)

Bachimont, Bruno. Antoine Isaac. Raphaël Troncy, Semantic Commitment for Designing Ontologies: A Proposal, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, p.114-121, October 01-04, 2002.

Um estudo de caso para aquisição de conhecimento no domínio da hematologia

Kátia C. Coelho[1], Mauricio B. Almeida[2,] Viviane Nogueira[3]

[1] Fundação Centro de Hematologia e Hemoterapia de Minas Gerais (Hemominas)
R. Grão Pará, 882 – Santa Efigênia – 30130110 – Belo Horizonte – Brasil

[2, 3] Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos, 6627 – Campus Pampulha – 31.270-901 – Belo Horizonte – Brasil.

katia.cardoso@hemominas.mg.gov.br, mba@eci.ufmg.br, vivianenpo@yahoo.com.br

Abstract. *Acquiring knowledge from experts has been a challenge within several research fields. There is no difference in the scope of Biomedicine, mainly because of the large amount of data. In this paper, we present an underway research being conducted in the domain of hematology, which investigates problems of the knowledge acquisition activity. A list of topics to carry out the knowledge acquisition was developed, which involve steps being applied in real world situations. The record of problems faced along the process aim to reach improvements in the best practices of the knowledge acquisition activity. In addition, we present and discuss partial results.*

Resumo. *Obter conhecimento especializado a partir de especialistas tem sido um desafio em diversos campos de pesquisa. Isso é especialmente verdadeiro em Biomedicina em função do volume de dados produzidos. No presente artigo, descreve-se pesquisa em andamento sobre dificuldades para aquisição do conhecimento no domínio da hematologia. Um roteiro para aquisição de conhecimento foi desenvolvido e tem sido aplicado empiricamente. Os registros das dificuldades ao longo do processo objetivam propor melhores práticas para a atividade. Resultados parciais são relatados e discutidos.*

1. Introdução

Investigar como o conhecimento produzido é traduzido em conhecimento de um campo científico, bem como os meios para organizá-lo e representá-lo é atividade essencial no âmbito da pesquisa. Entretanto, à medida que o volume de informação tem aumentado significativamente nos últimos 30 anos, representá-lo para uso por pessoas e por sistemas tem se tornando tarefa complexa.

Nesse contexto, ontologias têm sido propostas como alternativa para criação de representações da realidade. Desde os anos de 1980 têm sido estudadas como forma de representar conhecimento [Guarino, 1998], como referência para a criação de modelos [Fonseca, 2007] e para a representação do conhecimento [Vickery, 1997]. De forma simples, ontologias consistem de termos, relações e regras que regulam a combinação entre os termos organizados em uma taxonomia.

Ontologias são também definidas a partir do conjunto de processos que compõem a respectiva atividade de desenvolvimento [Fernandez, Gomez-Perez and Juristo, 1997],

como por exemplo, a aquisição de conhecimento. Trata-se de uma atividade sabidamente complexa em função das dificuldades de comunicação entre o engenheiro do conhecimento e o especialista [Milton et al 2006].

O presente trabalho se insere nesse contexto, enfocando a atividade de Aquisição de Conhecimento (AC) a partir de especialistas para o desenvolvimento de ontologias, conduzida no escopo de projeto em andamento no âmbito da biomedicina. Descreve-se a pesquisa que busca melhores práticas em atividades de AC, a partir da observação de dificuldades na obtenção de conhecimento de pesquisadores na Hemominas, instituição responsável pelo sistema hematológico e hemoterápico de Minas Gerais. A pesquisa busca, em última instância, estudar soluções para problemas como: a distância entre o conhecimento do especialista e o responsável pela AC - daqui em diante, "engenheiro" e barreiras entre especialista e engenheiro, como dificuldades de se obter consenso. O presente artigo apresenta o estágio atual da pesquisa: a metodologia utilizada para AC e as observações realizadas ao longo do processo na busca por respostas às perguntas citadas.

O restante do presente trabalho está organizado como segue: a seção 2 apresenta uma visão sobre AC. Cabe destacar que, por limitações de espaço, muitas questões são apenas pontuadas. A seção 3 apresenta a metodologia de pesquisa e a seção 4 discute resultados parciais. Finalmente, a seção 5 traz considerações e perspectivas futuras.

2. Aquisição de Conhecimento

A expressão AC é empregada desde a década de 1980 para se referir ao estudo da obtenção de *expertise* para representação em sistemas especialistas [Boose and Gaines, 1989] [Milton *et al*, 2006]. Diversas definições para AC são encontradas, mas é consenso que a atividade inclui pelo menos etapas de coleta, análise, estruturação e validação do conhecimento com finalidade de representação [Shadbolt and Burton, 1990].

A AC compreende um conjunto de tarefas que empregam teorias e métodos provenientes de campos diversos. Dentre essas disciplinas, cabe destacar a Ciência da Computação [Newell and Simon, 1975]; a Ciência Cognitiva [Hawkins, 1983]; a Linguística [Zellig Harris, 1976]; a Semiótica [Campbell, 1998] e a Psicologia [Kelly, 1955]. Cada um desses campos tem contribuído para a compreensão da atividade de AC.

2.1. Aquisição de Conhecimento: uma proposta de classificação das técnicas

Não há consenso sobre a classificação das técnicas de AC e diferentes propostas são encontradas na literatura, a saber: i) sob o ponto de vista do instrumento de aplicação: manual ou baseadas em computador (automáticas e interativas) [Boose and Gaines, 1989]; ii) em relação ao tipo de conhecimento obtido: procedural, conceitual, explícito, [Shadbolt, 2005] [Milton et al., 2006]; iii) do ponto de vista dos métodos [Shadbolt and Swallow, 1993]; e iv) dos métodos aplicados à biomedicina [Payne et al, 2007], esses últimos, relevantes no contexto desse trabalho.

As técnicas manuais, como a Grade de Repertórios, têm raiz na Psicologia [Boose and Gaines, 1989]. As técnicas interativas fazem uso de algum tipo de ferramenta para a interação do engenheiro com o especialista. Exemplos de técnicas automatizadas, como

o aprendizado da máquina, são apresentados no Projeto Neon[1] [Maynard and Nioche, 2006] [Gomez-Perez, Erdmann and Greaves, 2007]

As técnicas para AC orientadas para os métodos de aplicação podem variar de acordo com o tipo de conhecimento que se objetiva eliciar. Exemplos de tais técnicas são: a) geração de protocolos, entrevistas e observação; b) *técnicas baseadas em matriz; c) técnicas de ordenamento e;* d) *técnicas de limitação e restrição de informação* [Shadbolt and Swallow, 1993].

2.2. Aquisição de conhecimento em Biomedicina

A organização de terminologias na área biomédica é um desafio constante, em função da amplitude do assunto e da multiplicidade de interpretações para os dados e suas formas de obtê-los [Smith, 2008]. A literatura AC em biomedicina apresenta diferentes propostas para representação do conhecimento médico e biológico.

Vita et al (2006) descrevem a curadoria, parte do processo de anotação automatizado, quase uma exigência face à quantidade de artigos científicos produzidos. As anotações geradas exigem análise e validação por especialistas. Um exemplo é a extração de dados imunológicos que exigem alto nível de especialização. Stehr et al (2010) descrevem curadoria em redes colaborativas. Hoehndorf et al (2009) também se valem de uma interface *wiki* baseada em ontologias para aquisição semi-automática de conhecimento.

3. Metodologia de pesquisa

Os entrevistados para AC são especialistas do grupo de pesquisa sobre HTLV (vírus linfotrópico de células T humanas) em Minas Gerais. Um roteiro experimental para AC foi proposto, como forma de realizar a pesquisa fim: identificar dificuldades ao longo da atividade de AC em biomedicina. O roteiro foi elaborado a partir da revisão de literatura. Ao realizar uma AC real, o pesquisador vem registrando problemas de forma a propor melhorias. O roteiro de tarefas para AC contempla três fases: i) Levantamentos; ii) Contatos; 3) Validação.

A *fase de levantamentos* compreende atividades anteriores ao contato direto com especialistas. A primeira tarefa é conhecer o escopo da ontologia de domínio, conhecer a que fim a se destina, quem vai utilizá-la e com que propósitos, além de dados sobre a ontologia de alto nível. Além disso, é necessário identificar os especialistas, bem como o conhecimento que produzem e registram em fontes. Exemplos de levantamento dos especialistas e *expertise* são apresentados na Tabela 1.

Tabela 1. Extrato do levantamento de especialistas para atividade de AC

Experts [(*)]	**A. B.C.F.**	**M.A R.**	**D.U.G.**	**E.F.B.**	**M.S.N.**
Formação	Médico	Médico	Médico	Veterinário	Médico
Atuação	Pesquisa	Pesquisa e Medicina	Pesquisa e Ensino	Pesquisa e ensino	Pesquisa
Linhas de Pesquisa	Epidemiologia	Epidemiologia	Otoneurologia	-	Clinica médica
	Hematologia	Infectologia	Infectologia	-	Hematologia
	Virologia	-	-	Virologia	-

[(*)] os especialistas são aqui identificados por códigos

[1] Disponível na internet em: http://www.neon-project.org/ Acesso: 21 julho de 2011

A *fase de contatos* consiste no encontro com especialistas e realização de atividades que permitam obter e registrar conhecimento. Exemplos de técnicas utilizadas são: i) entrevistas, baseadas em *template* no *Protegé-Frames* baseado em Scheuermann et al (2009); ii) técnicas baseadas em matriz; iii) técnica de ordenamento,. A *fase de validação* faz uso de ferramentas *wiki* para colaboração na análise de termos candidatos a ontologia, bem como de suas definições. A partir do conhecimento obtido na AC, termos candidatos a ontologia são transpostos para uma wiki, onde em seguida são validados pelos especialistas via internet.

Tabela 2. roteiro de AC utilizado na pesquisa

Fase	Objetivo da tarefa	Descrição da Tarefa	Instrumento
(1) Levantamento	1.1 conhecer o contexto	Conhecer escopo da ontologia em desenvolvimento	Dados do projeto
	1.2 conhecer fundamentos	Conhecer conceitos básicos do domínio em questão	Literatura básica da área
	1.3 identificar *expertise*	Identificar *expertise* dos especialistas envolvidos	Diretórios de pesquisadores
(2) Contato	2.1 obter conhecimento	Entrevistas realizadas com especialistas	*Template Protege-Frames*
	2.2 conhecer terminologia	Identificar problemas para organização da informação	Técnicas de Matriz
	2.3 ver organização *ad-hoc*	Entender como especialistas ordenam conceitos	Técnicas de ordenamento
(3) Validação	3.1 validar conhecimento	Obter aprovação sobre termos adquiridos e suas definições	Página Wiki
	3.2 manter conhecimento	Atualizar dados após cada validação	Página Wiki

4. Resultados parciais

Conforme já mencionado, a pesquisa tem sido realizada no contexto de projeto de organização da informação sobre biomedicina. Ao longo da atividade de AC têm sido observadas e registradas problemas relativos aos especialistas e aos engenheiros. As principais observações são descritas brevemente à seguir. Os resultados parciais foram obtidos através de observação ao processo de AC em questão e, no atual estágio de pesquisa, não representam amostra quantitativa significativa.

Especialistas:

- O especialista tem pouco tempo disponível por acumular outras atividades como: ensino, coordenação, atendimento clínico entre outras;
- O especialista tem dificuldade de explicitar o que sabe e desconhece ontologias;
- Existe super-especialização no âmbito da própria especialidade, o que resulta em narrativas com diferentes níveis de granularidade e formas de organizar a informação;
- O especialista sugere fontes adicionais de conhecimento como complemento ao seu relato na AC, tendo em vista que suas inúmeras publicações;
- O uso de entidades provenientes de ontologias de alto nível, de forma a apresentar a organização hierárquica preliminar, acaba por confundir o especialista;

Engenheiro

- A curva de aprendizado no domínio da biomedicina é longa e complexa;
- O vocabulário especializado dificulta a compreensão das terminologias;
- O assunto é multidisciplinar e os especialistas atuam em sub-domínios: existem poucos especialistas para consulta e eles são super-especializados;
- Entidades de provenientes do alto nível usadas para orientar o engenheiro, nem sempre encontram respaldo no dia a dia do especialista, o que dificulta o diálogo.

As observações registradas até o momento resultam em recomendações: a) o uso de fermenta de apoio como *Protegé-Frames* para organizar as entrevistas de acordo com princípios ontológicos sistemáticos; b) a consideração de AC semi-automática pelo grande volume de publicações; c) a imersão da literatura é acompanhada de técnicas como análise de assunto, análise conceitual, dentre outras; d) a realização da AC tem lugar ao longo das atividades rotineiras do especialista; e) uso de ferramentas interativas para que o especialista registre, ele mesmo, o que sabe, sem intervenção do engenheiro.

5. Considerações finais

Este artigo apresentou pesquisa em andamento que busca por melhorias nas práticas de AC. Para tal, descreveu-se literatura sobre AC (inclusive no âmbito da biomedicina), apresentou-se um roteiro para a AC, e discutiu-se resultados parciais através da identificação de dificuldades na comunicação.

A continuidade da pesquisa abordará questões ainda em aberto e buscará a confirmação de resultados parciais ,na busca por respostas à questões mencionadas na introdução do presente artigo, a saber: como lidar com barreiras no processo de AC como: a formação de consenso; o desconhecimento sobre o trabalho do engenheiro; a falta de tempo do especialista? Que favorecem o consenso em ambientes colaborativos? Como lidar com diferenças terminológicas entre os próprios especialistas em ambientes colaborativos?

Agradecimentos

Este trabalho conta com apoio da Fundação Hemominas–e da FAPEMIG

Referências

Boose, J. H., and Gaines, B. R. (1989) “Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the-Art”. http://www.springerlink.com/

Campbell, K. E. et al. (1998) “Representing thoughts, words, and things in the UMLS”. http://www.ncbi.nlm.nih.gov/pubmed/9760390, October.

Fernandez, M., Gomez-Perez, A. and Juristo, H. (1997) “Methontology: From Ontological Art Towards Ontological Engineering”, http://www.aaai.org/Papers/Symposia/Spring/1997/SS-97-06/SS97-06-005.pdf

Fonseca, F. (2007) “The Double Role of Ontologies in Information Science Research”. http://citeseerx.ist.psu.edu/, April.

Harris, Z. (1976) “On a theory of Language”. http://www.jstor.org/stable/2025530, May.

Hawkins, D. (1983) “An analysis of expert thinking”. http://www.sciencedirect.com/ science/ journal/00207373, January.

Hoehndorf, R. et al. (2009) “BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology”. http://www.biomedcentral.com/1471-2105/10/S5/S5, July.

Hoffman, R.R. (1995) “Eliciting Knowledge from Experts: A methodological analysis”. Organizational Behavioral and Human Decision Processes. Vol. 62, no 2, p. 129-158.

Gomez-Perez, J.M., Erdmann, M. and Greaves, M. (2007). “Applying problem solving methods for process knowledge acquisition, representation, and reasoning”. Proceedings of the K-CAP 2007

Kelly, G. (1955). Princípios da Psicologia dos Construtos Pessoais. Norton, 3rd edition.

Maynard, D. and Nioche, J. (2006) “Human Language Technology for Knowledge Acquisition for the Semantic Web”. EKAW 2006

Milton, N., Clarke, D. and Shadbolt, N. (2006) “Knowledge engineering and psychology. http://portal.acm.org/citation.cfm?id=1221471, December.

Newell, A. and Simon, H.A. (1975) “Computer science as empirical inquiry: symbols and search” http://citeseerx.ist.psu.edu/, March.

Nilsson, N. (2007) “The Physical Symbol System Hypothesis: Status and Prospects”. http://ai.stanford.edu/~nilsson/OnlinePubs-Nils/PublishedPapers/pssh.pdf

Scheuermann, R. et al. (2009) “Toward an ontological treatment of diasease and diagnosis”. http://ontology.buffalo.edu/medo/Disease_and_Diagnosis.pdf, February.

Shadbolt, N. (2005) “Eliciting Expertise. http://eprints.ecs.soton.ac.uk/14563/1/Elciting_Expertise.pdf,

Smith, B. (2008). New Desiderata for Biomedical Terminologies. In *Applied Ontology*, pages 21–39. Ontos-Verlag.

Stehr, H. et al. (2010) “PDBWiki: added value through community annotation of the Protein Data Bank”. http://database.oxfordjournals.org/content/2010/baq009.full, March.

Vickery, B. C. (1986) “Knowledge representation: a brief review”. http://www.emeraldinsight.com/, September.

Vita, R et al. (2006) “Curation of complex, context-dependent immunological data”. http://www.biomedcentral.com/1471-2105/7/341, July.

Wolf, R. and Delugach, H. S. (1996) “Knowledge Acquisition via tracked repertory grids”. http://www.cs.uah.edu/tech-reports/TR-UAH-CS-1996-02.pdf

Desenvolvimento de Ontologias para o Portal Semântico do CPDOC

Renato Rocha Souza[1], Suemi Higuchi[2], Daniela Lucas da Silva[3]

[1]Escola de Matemática Aplicada – Fundação Getúlio Vargas (EMAp – FGV).

[2] Centro de Pesquisa e Documentação de História Contemporânea do Brasil – Fundação Getúlio Vargas (CPDOC – FGV).

[3]Departamento de Biblioteconomia – Universidade Federal do Espírito Santo (UFES).

{renato.souza,suemi.higuchi}@fgv.br, danielalucas@hotmail.com

Abstract. *This paper describes the semantic portal project being developed at CPDOC – FGV, along with all the initiatives that are being undertaken in order to achieve the final goal. Among those initiatives we can highlight the domain ontology creation, in the field of Brazil's contemporary history and document description, for the proper metadata supply to the documents from the archives of interest.*

Resumo. *Este artigo descreve o projeto de criação do portal semântico do CPDOC – FGV, juntamente a todas as iniciativas que estão sendo engendradas para que este seja possível. Dentre estas, destacam-se a criação de ontologias de domínio para história contemporânea e descrição de acervos, para o adequado provisionamento de metadados para os documentos pertencentes aos acervos em questão..*

1. Introdução

O Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) é parte da Escola de Ciências Sociais e História da Fundação Getúlio Vargas. Criado em 1973, tem o objetivo de abrigar conjuntos documentais relevantes para a história recente do país, desenvolver pesquisas históricas e promover cursos de graduação e pós-graduação. Dentre estes conjuntos documentais, podem-se distinguir aqueles doados por importantes personalidades da história brasileira contemporânea e aqueles que são fruto de atividades de pesquisas do próprio CPDOC, como entrevistas, dossiês e dicionários de verbetes. Os conjuntos documentais são organizados em sistemas, com características próprias, como detalhado a seguir:

- Arquivos pessoais, doados ao CPDOC como partes do espólio de personalidades públicas constituem, atualmente, o mais importante acervo de arquivos pessoais de homens públicos do país, integrado por aproximadamente 200 fundos, totalizando cerca de 1,8 milhão de documentos, entre textos, imagens e vídeos.
- Programa de História Oral do CPDOC, que desde 1975 vem produzindo um acervo de depoimentos (em áudio e vídeo) de importância reconhecida tanto no Brasil como no exterior. No total são cerca de 1.000 entrevistas, correspondendo a mais de 5 mil horas de gravação, estando metade delas abertas à consulta na web.

- Dicionário Histórico Biográfico Brasileiro (DHBB). Começou a ser desenvolvido no CPDOC-FGV em 1974 e gerou uma primeira versão impressa em quatro volumes com 4.493 verbetes sobre conceitos da História Contemporânea Brasileira. Lançada em 2001, a segunda edição do DHBB em formato de CD-ROM atualizou os verbetes existentes e incluiu novos, atingindo um total de 6.620 entradas. A versão atual, lançada em 2010 na web, compreende 7.553 verbetes, sendo 6.584 de natureza biográfica e 969 verbetes temáticos, relativos a instituições, eventos e conceitos de interesse para a história do Brasil pós-1930.

Em 2008 o CPDOC iniciou um amplo projeto de digitalização do seu acervo, ainda em curso. Em 2010 o acervo digitalizado continha a conversão de mais de 300 mil documentos textuais, 65 rolos em película, 106 fitas (VHS, Beta e U-MATIC), 350 discos, 187 fitas cassete, 85 fitas rolo e cerca de 32.000 fotografias do acervo de Arquivos Pessoais. Além disso, foram digitalizadas 5.000 horas de entrevistas do Programa de História Oral, estando toda esta documentação disponível para consulta no CPDOC . Ao final do projeto, conta-se com cerca de 80.000 fotografias digitalizadas disponíveis para consulta através da web, dando conta de praticamente todo o acervo de imagens doado até 2010 para o Centro. Além disso, todos os verbetes do DHBB se encontram em formato digital.

A característica comum aos acervos reside no fato de conterem documentos em mídias diversificadas, como texto manuscrito, texto em formato digital, áudio com e sem transcrições, imagens e vídeos com e sem legendas, caracterizando a multimodalidade midiática que apresenta difícil tratamento para fins de recuperação, e a publicização destes acervos vem sendo realizada através de interfaces e processos distintos, apesar de serem abrigados por uma única instituição e poderem ser acessados através do mesmo portal.

Em 2008, foi criada na FGV a Escola de Matemática Aplicada (EMAp), tendo como missão atuar na aquisição e repasse do conhecimento científico e tecnológico de base matemática para utilização nas áreas de interesse da FGV e parceiros. Em contato inicial com o CPDOC, propôs-se uma parceria para aplicação das técnicas de recuperação de informação desenvolvidas no escopo da Matemática Aplicada para uso no CPDOC. A partir deste contato, foi realizado um diagnóstico nos sistemas de informação do CPDOC que apontava, de maneira geral, para a necessidade de maior integração entre os sistemas e melhoria na descrição dos dados e nas interfaces de acesso. Estes motivadores levaram à criação de projetos de parceria que, em termos gerais, buscam melhorar a integração e gestão dos sistemas de informação, e acesso externo aos acervos, aumentando a visibilidade dos arquivos salvaguardados e das produções intelectuais desenvolvidas para a sociedade.

Neste artigo descreve-se em linhas gerais o projeto do portal semântico do CPDOC, e especificamente sua vertente que envolve o desenvolvimento de ontologias. O projeto prevê a migração de todo o acervo atual para uma base de dados comum em formato *RDF triplestore*, e a unificação dos padrões de descrição entre todos os fundos e sistemas, o que envolve a criação de ontologias de descrição e de domínio. Como objetivo, pretende-se oferecer uma interface única para buscas temáticas transversais e integradas, utilizando-se conceitos e categorias de conceitos relativos ao domínio da

História Contemporânea Brasileira – como pessoas, acontecimentos e locais – através de todos os sistemas/acervos atuais.

2. O problema

O principal problema a ser enfrentado se caracteriza pelo tratamento integrado de bases heterogêneas e em formatos multimídia, e a ausência de padronização nos formatos de descrição. No âmbito do projeto, almeja-se uma interface única e um padrão unificado de metadados para descrição dos inúmeros itens dos diversos acervos.

Como foram construídos de maneira independente, os acervos, sistemas e fundos adotaram padrões idiossincráticos de descrição, ressaltando diferentes características a serem descritas e diferentes terminologias para descrevê-las. Acrescenta-se a esta dificuldade o fato de ser o acervo composto por fotografias, cartas, desenhos, periódicos, entrevistas em áudio e vídeo, gravações de rádio, de vídeo, dentre outros.

3. A solução proposta

O problema proposto demanda uma série de iniciativas razoavelmente independentes de preparação dos acervos e sistemas para a migração. Estas iniciativas são descritas a seguir:

- Projeto de reconhecimento de faces e personagens: teve como objetivo otimizar os processos de gestão do acervo fotográfico do CPDOC, a partir de técnicas de reconhecimento de faces e de personagens. Como resultado, foram desenvolvidos aplicativos para tratar os fundos organizados com legendas, realizando a detecção de faces e a combinação destas com as legendas já produzidas. Além disso, atende à demanda do CPDOC de disponibilizar ao público de maneira mais amigável a localização dos personagens em cada fotografia de nosso acervo.
- Projeto de alinhamento de som e texto: teve como objetivo produzir transcrições automáticas de voz em língua portuguesa, a serem utilizadas pelo o programa de história oral do CPDOC no tratamento de seus acervos. O material utilizado é constituído de entrevistas transcritas, entrevistas gravadas – arquivo de áudio, transcrições das entrevistas – arquivo de texto, entrevistas sem transcrição, entrevistas gravadas – arquivo de áudio, Sumário das entrevistas – arquivo de texto.
- Projeto de mineração de textos: é, na verdade, um conjunto de iniciativas de processamento de linguagem natural para oferecer, entre outras coisas, Suporte aos projetos reconhecimento de faces e personagens e de alinhamento de som e texto. Nesta iniciativa, foram coletados possíveis descritores (termos frequentes encontrados em legendas de fotos, em documentos, e em transcrições de entrevistas) com vistas à incorporação nas ontologias de domínio e também no DHBB.
- Projeto de “Wikificação” do DHBB: Foi engendrado para promover uma maior interligação das bases de dados internas do CPDOC com as externas, como a própria Wikipédia, com benefícios no sentido de aumento da publicização e estruturação de redes sociais de colaboração para contribuições e eventuais correções para o acervo. Está sendo implementada através de uma ferramenta open source de Wiki Semântico (MediaWiki com extensões semânticas), e nesta *wiki* estão sendo cadastrados verbetes do DHBB para demonstrar as possíveis funcionalidades do ambiente. Este projeto se beneficia das ontologias que estão sendo criadas.

- Projeto de Criação de Ontologia a partir dos Descritores de Sistemas: a descrição dos acervos do CPDOC é realizado hoje através de uma enorme lista não hierárquica de descritores, que contém, entre outras coisas, instâncias de pessoas, entidades, processos, eventos, locais e atributos. Esta lista se constitui no primeiro levantamento de conceitos para a criação da ontologia de história contemporânea, junto aos verbetes hoje presentes no DHBB.

4. O portal semântico toma forma

Todos estes projetos são fins em si, com utilidade e potencial de melhorias imediatas para os sistemas como se encontram atualmente. Mas a culminação dos projetos constitui o embrião do Portal Semântico do CPDOC. Este compreende uma solução de migração dos acervos para um sistema único, com tecnologias abertas e preconizadas pelo W3C. A proposta de Portal encerra uma solução que proporcionará:

- Acesso unificado aos acervos dos sistemas;
- Navegação e busca pautada por conceitos, independente de mídias e de sistemas;
- Buscas transversais entre sistemas (DHBB, Arquivos Pessoais, PHO, etc.);
- Interligação dos acervos através de conceitos comuns;
- Padrões únicos de descrição de itens entre os sistemas;
- Padrões de descrição adotados mundialmente, permitindo a interoperabilidade e interligação com sistemas e acervos externos;
- Integração com os repositórios da web (*Linked Data / Linked Open Data*[1]) através da utilização de uma base de dados em padrão *RDF triplestore*;
- Conceitos relevantes estruturados sob a forma de verbetes, com nome e endereço únicos, preferencialmente sob a forma de URIs;
- Maior visibilidade do acervo sob a ótica dos mecanismos de busca;
- Possibilidades aumentadas de integração dos acervos como objetos educacionais;
- Possibilidade de integração com a Biblioteca Digital da FGV;

Dentre outros aspectos. A FIG.1. A seguir exemplifica o esquema do Portal Semântico com os processos de conversão de bases:

[1] http://linkeddata.org/

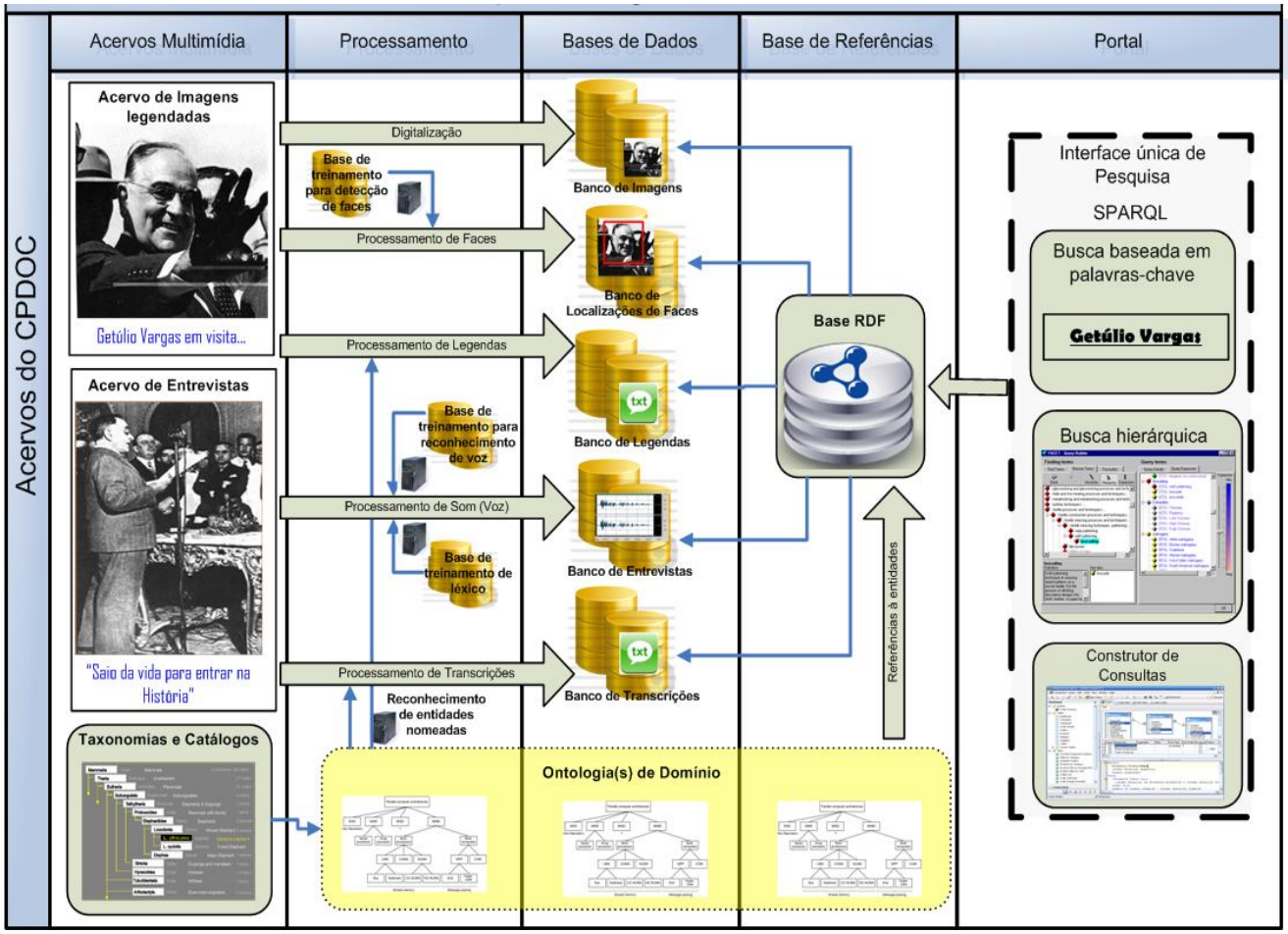


Figure 1. Esquema do Portal Semântico

Para materializar a solução integrada, em conjunto e adicionalmente aos projetos apresentados, serão desenvolvidas as seguintes ações de nível macro:

- Análise dos metadados descritivos de cada fundo/acervo;
- Criação de um formato padronizado de descrição, preferencialmente através da composição de ontologias existentes e utilizadas no escopo da web semântica (Dublin Core, Bibliographic Ontology, FOAF, dentre outros), e que sejam compatíveis com os padrões de descrição arquivística (NOBRADE, ISAAR(CPF), ISAD(G));
- Criação de Ontologias leves (light ontologies) no domínio da História Contemporânea para descrição do conteúdo dos documentos;
- Classificação dos documentos segundo os campos dos padrões de descrição adotados, e utilizando os conceitos desenvolvidos na Ontologia de História Contemporânea;
- Classificação das fotografias através da digitalização e processamento de legendas, e através de técnicas de reconhecimento de faces e de personagens;

- Classificação do material em áudio, através de processamento de transcrições e análise dos campos de metadados;
- Migração dos acervos para uma nova base de dados em formato triplestore, ou seja, um banco de dados próprio para armazenamento de dados no formato RDF;
- Interligação interna e externa dos itens dos acervos através de identificadores únicos, preferencialmente acessíveis via hipertexto (URIs e URLs);
- Criação de interfaces de pesquisa no acervo através da tecnologia SPARQL.

As ontologias a serem criadas – de descrição e no domínio da história contemporânea – serão desenvolvidas segundo a metodologia híbrida proposta em Silva, Souza e Almeida (2008), dando-se prioridade ao reuso de ontologias existentes, no caso específico das ontologias de descrição bibliográfica.

5. Discussão

Este artigo apresenta o panorama de um projeto que se encontra em plena execução, tendo sido iniciado em 2010 e com previsão de término para o final de 2012. Envolve uma série de iniciativas que estão sendo desenvolvidas em paralelo, com um horizonte de unificação, materializada na conjuminação de cinco projetos independentes, como foi apresentado, além de ações específicas do projeto do portal semântico. Constitui um projeto representativo de recuperação de informações multimodal porque lida com documentos em formatos diversificados, como áudio, vídeo, imagens, textos e modelos conceituais. Além disso, incorpora tecnologias e instrumentos oriundos do ferramental da web semântica, como triplestores RDF e ontologias. O produto final, acredita-se aumentará enormemente a publicização e acesso dos acervos e sistemas mantidos pelo CPDOC, contribuindo para seu melhor uso pela sociedade em geral.

6. Referências

SILVA, Daniela Lucas da ; SOUZA, Renato Rocha ; ALMEIDA, Maurício Barcellos de. Ontologias e vocabulários controlados: comparação de metodologias para construção. Ciência da Informação (Impresso), v. 37, p. 60-75, 2008.

Ontology Merging: on the confluence between theoretical and pragmatic approaches

Raphael Cóbe[*], Renata Wassermann, Fabio Kon

[1]Department of Computer Science – University of São Paulo (IME-USP)

{rmcobe,renata,fabio.kon}@ime.usp.br

Abstract. *In recent years, researchers have focused on merging knowledge bases in both pragmatic and theoretical points of view. In this paper, we enumerate a few attempts to deal with inconsistencies while merging knowledge bases. We focus on ontology merging and show that pragmatic and theoretical approaches are not integrated and that both could benefit from a closer relationship. We extended an existing theoretical algorithm for Description Logics and applied it for the ontology merging problem. We describe here an implementation of this algorithm as an open source Protégé plugin.*

1. Introduction

There has been a rapid increase in availability of (semantic) information on the web. Nevertheless, there is no standard way of reusing knowledge, creating a challenge of building new knowledge bases for specific domains. This has forced users to build knowledge bases from scratch instead of being able to reuse previously established knowledge.

Ontologies have been considered as a mean for expressing and sharing semantic knowledge among systems [Gruber 1993] specially in the context of the Semantic Web. Their underlying structures allow machine-processing, providing a common vocabulary for expressing metadata about each web resource. Also, they are based on first order logic, allowing the usage of reasoners that are able to infer relationships between concepts based on their logical description. In that sense, W3C proposed the OWL[1] standard specification language to express ontologies.

The main challenge faced by knowledge integration research is solving inconsistencies by removing the minimum amount of information so that the remaining stays consistent. Since we are talking about removing part of the knowledge from the base, it is important to clarify the difference among three kinds of knowledge integration: (a) merging, (b) revising, and (c) updating. Revesz [Revesz 1995] claims that revision and update operators are characterized by the inclusion of knowledge that is either more or less relevant (or trustful) than the knowledge previously defined, while merge, in contrary, does not prefer any piece of knowledge over another.

That kind of concern resulted in works like [Konieczny and Pérez 1999] where the authors have dealt with the merging problem for the propositional logic case by means of a model-based approach. This kind of work has inspired most of the research conducted in the theoretical field of first order logic merging, such as the work of Gorogiannis et al. [Gorogiannis and Hunter 2008] and Qi et al. [Qi et al. 2006]. Unfortunately, pragmatic approaches did not follow the same evolution pace as theoretical ones and just a

[*]The author would like to thank FAPESP for sponsoring his research.

[1]www.w3.org/TR/owl-features/

few tools have been developed to provide knowledge base integration. This might have happened because pragmatic research has focused on the ontology mapping activity. If we think about the whole knowledge integration as a process: first of all, we have to compute whether there are similar concepts and how similar such concepts are - this is the *mapping activity* - and each concept correspondence is called concept *match*. After mapping the concepts, the merging activity takes place. During the merging, the concepts from all the knowledge bases are copied into the output base. Thus, mapping is the activity that most of the time comes before the merging (at least in the ontology integration) [Falconer et al. 2007].

In this paper, we present our current work on ontology merging, including the implementation of a plugin for the Protégé[2] editor.

This paper is organized as fallows. Section 2 presents a brief summary of pragmatic and theoretical works that aim to deal with ontology inconsistencies. Section 3 explains our efforts in bridging the gap between both points of view and present the merging plugin we developed. Finally, at Section 4, we present a few conclusions taken from our work and discuss what we plan to do in the future.

2. Ontology Merging

In this section, we intend to show the common approaches used to deal with the ontology merging and inconsistency handling problem. We have divided this section in two to show that these two fields of study are not dealing with the same problems.

2.1. Theoretical Approaches

Only a few studies in the literature deal directly with description logics based knowledge integration and inconsistency management. We classify these works, like [van Harmelen et al. 2005], into two main categories: syntactic and semantic-based approaches. The syntactic-based approaches sees ontologies as a set of axioms, which are syntactic objects, while semantic-based approaches sees ontologies as a set of models, which are semantic objects that are represented by a finite set of axioms.

In the context of syntactic-based approaches, we would like to cite the research conducted by Thomas Meyer and his colleagues. They proposed an algorithm for finding maximally consistent sets from inconsistent knowledge bases [Meyer et al. 2005]. This algorithm is a modification of the *conjunctive maxi-adjustment* algorithm for propositional knowledge integration and is called *CMA-DL*. In that sense, such work is similar to the work developed by van Harmelen et al. [van Harmelen et al. 2005], but instead of looking for maximally consistent subsets their goal is to build minimally inconsistent subsets, which they call diagnoses. The main difference between these two approaches is that the CMA-DL algorithm takes into account the order of the bases to be merged. Each base is called *strata* and the set of all strata is called *stratified knowledge base*. This set is ordered by preference, which means that the first ontology is preferred over the second one during the merging activity. We have proposed a small modification to this algorithm that gives to the user all possible merging precedence order. We have used this algorithm to implement our Protégé merging plugin.

Most of the semantic-based approaches have been inspired by model-based propositional logic inconsistency solving like what is presented at [Konieczny and Pérez 1999].

[2] `http://protege.stanford.edu`

In that context, Gorogiannis et al. [Gorogiannis and Hunter 2008] propose an approach to deal with inconsistencies by means of *Dilation Operators* that are, basically, an strategy to iteratively relax the formulas to remove inconsistencies. The authors have first proposed the use of Dilation Operators to deal with inconsistencies in propositional logics and showed the equivalence of their approach to the one from Konieczny and Pérez [Konieczny and Pérez 1999]. Finally, they took the idea of using Dilation Operators further and proposed an operator that iteratively transform first order formulas by changing universal quantifiers into existential ones. This approach may solve a few inconsistencies but we figured out that it would be hard to translate it to an ontology context. For instance, the description logic formula equivalent to $\forall x.p(x) \rightarrow z(x)$ would be $p \sqsubseteq z$ but we were unable to define a way to dilate the description logic formula so it would be equivalent to the dilated first order logic formula, i.e. $\exists x.p(x) \rightarrow z(x)$. Qi and colleagues have also proposed model-based operations to solve inconsistencies in ontologies. In [Qi et al. 2006] they proposed a model-based operator named *weakening* and showed that its results are semantically equivalent to those of CMA in stratified knowledge bases.

2.2. Pragmatic Approaches

In this section, we present a few tools which purpose is to manage multiple ontologies to combine and promote the reuse of knowledge. We will discuss the PROMPT approach for ontology merging and specially inconsistency handling in more detail as it was the only one that we have found that deals with inconsistency. We have tried a few other tools but, unfortunately, none of them provided any inconsistency handling mechanism. For instance, we have tried Watson For Knowledge Reuse[3], which is a tool that allows the user to query a web service that contains ontologies and ask it for suggestions on new concepts to be added. It may suggest to include relationships and concepts that may break the ontology consistency. So, it is not specially concerned with keeping the ontology consistency. We have also tried the OWLDiff[4] tool. It intends to work just like the common Unix *diff* command, providing an easy-to-use interface that shows the differences between two ontologies. It also allows the user to copy ontology fragments between ontologies but does no consistency checking after doing so.

We have studied the PROMPT tool for merging and it does provide support for inconsistency management. Unfortunately, the inconsistencies dealt with PROMPT strategies are not logic inconsistencies and they arise due to the fact that its merging algorithm sometimes fail in merging concepts and properties. We figure that PROMPT does not deal with logical inconsistencies because when the authors proposed its idea [Noy and Musen 2000], the Protégé OWL tool did not provide support for more expressive logics constructions like the disjoint clause.

PROMPT deals with 4 kinds of inconsistencies: a) Name Conflict: this inconsistency happens when the algorithm includes two different concepts with the same name in the merged ontology, so the system advises the user to rename them; b) Dangling References: this inconsistency happens when the image of a given property is missing in the merged ontology, so the system suggests that the user includes such concept into the merged ontology; c) Redundant Hierarchy: this inconsistency happens when there is more than one path connecting a concept to one of its ancestors, so the system suggests that the

[3] `http://neon-toolkit.org/wiki/1.x/Watson_for_Knowledge_Reuse`
[4] `http://krizik.felk.cvut.cz/km/owldiff/`

user removes one of these paths; d) Slot Constraint Violation: this inconsistency happens when some property has its cardinality violated at the merged ontology, e.g. a property that should have only one individual as its images is used to connect two different pair of individuals. The systems then suggests that the user removes one of these individuals.

3. On the confluence of theoretical and pragmatic approaches

The main focus of the work that we are currently developing is to bridge the gap between the theoretical and pragmatic approaches for ontology integration. Unfortunately, pragmatic approaches have very little to offer since the only conflict solving approach described (by PROMPT) does not deal with logical inconsistencies. Also, the theoretical works most of the time cannot be directly applied to Ontology merging, their algorithms were designed for using with first order logics like [Gorogiannis and Hunter 2008].

We chose to use the CMA-DL algorithm designed by Meyer et al. [Meyer et al. 2005] to solve merging inconsistencies in description logics and applied it to ontology merging. The algorithm proposed the generation of maximally consistent subsets of the axioms present at each ontology at the inconsistent knowledge base in an iterative way. We chose to use such algorithm as the starting point for our research because it is a syntactic-based approach that can clearly build maximally consistent ontologies, not like the algorithm from [van Harmelen et al. 2005], which relies on a *Connectedness* notion and the proposed *Direct Structural Connection* function cannot detect axioms that cause inconsistencies that are not structurally connected.

The CMA-DL algorithm takes into consideration that, at the inconsistent knowledge base, the ontologies are sorted by order of preference. We believe that, sometimes, this is the case and we believe that this approach is close to the knowledge revision, but sometimes we cannot classify the ontologies according to their relevance. Let us take a look at the example (taken from [Meyer et al. 2005]) for instance:

Example 1: Consider the knowledge base $K = (S_1, S_2)$ and that the ontology S_1 is composed of the following axioms $bird(tweety), \neg flies(tweety), bird(chirpy)$ and that the ontology S_2 is composed of the axiom $bird \sqsubseteq flies$. It is easy to see that this knowledge base is inconsistent, sice S_1 states that *tweety* is a bird that cannot fly and S_2 states that every bird flies.

The CMA-DL algorithm gives preference to the knowledge present in S_1 and the result for its processing is the ontology O composed of the axioms exclusively from S_1, i.e., $O = \{bird(tweety), \neg flies(tweety), bird(chirpy)\}$. Although this is an ontology free of inconsistencies, we argue that at this case the discarded axiom seems to be very important to the result. It is a constraint that applies to all individuals of the bird concept.

We have proposed a modification to the CMA-DL algorithm that takes into account all possible ordering combinations of the input knowledge base. Such algorithm can be seen at the Listing 1. It accumulates the results of the possible combinations and leaves to the user the choice of which one to use. The algorithm presented relies on a set of specific operations. It uses a *PowerSet()* operation to calculate all possible subsets of a given set and *Permutations()* to calculate all possible permutations of a given set. It also relies on the *Ontology()* operation to generate a new Ontology from a set of axioms given and, conversely, it uses the *Axioms()* operation to retrieve a set containing all axioms of a given ontology.

Listing 1. Modified CMA-DL

```
1  Algorithm M-CMA-DL:
2    Input: A set of ontologies 𝒟 := (D_1, D_2, ..., D_n)
3    Output: A Set of Consistent Ontologies
4    ℬ := {An Empty Ontology}
5    for all Permutation (O_1, O_2, ..., O_n) in Permutations(𝒟) do
6      for i := 1 to n do
7        𝒞 := ℬ
8        for all Ontology C in 𝒞 do
9          𝒪 := Axioms(O_i)
10         j := ‖𝒪‖
11         𝒮 := ∅
12         repeat
13           𝒳 := {X ‖ X ∈ PowerSet(𝒪) and ‖X‖ = j}
14           for all Axiom Set X in 𝒳 do
15             if Ontology(Axioms(C) ∪ X) is consistent then
16               𝒮 := 𝒮 ∪ {X}
17             endif
18           endfor
19           j := j − 1
20         until 𝒮 is not empty
21         for all Axiom Set S in 𝒮 do
22           ℬ := (ℬ \ {C}) ∪ {Ontology(Axioms(C) ∪ S)}
23         endfor
24       endfor
25     endfor
26   endfor
27   return ℬ
```

If we run our version of the algorithm using as input the same knowledge base presented in Example 1, we would get the following set as output: $R = \{O_1, O_2, O_3\}$, where $O_1 = \{bird(tweety), \neg flies(tweety), bird(chirpy)\}$, $O_2 = \{bird(tweety), bird(chirpy), bird \sqsubseteq flies\}$, and $O_3 = \{bird(chirpy), bird \sqsubseteq flies\}$. One can easily see that our approach gives more power of choice to the user and at two different options he/she is able to keep the axiom that states that every bird flies, which was our primary goal.

We have developed a Protégé view plugin (Figure 1). that allows users to merge two ontologies at each time. The plugin can use both the classic and our version of the CMA-DL. It uses the HermiT[5] reasoner to check the consistency and OWLAPI[6] to access and manipulate ontologies. It is distributed under the GPL v3.0 license and is available at `http://ccsl.ime.usp.br/en/onair/ontology-merging`. The plugin allows the user to pick two ontologies from his/her filesystem and choose the destination where the resulting ontologies are going to be stored. Lastly, the user chooses whether he/she wants to use the classic CMA-DL or our modified version by checking the option "Merge ontologies using the first one as more important (Classic CMA-DL)".

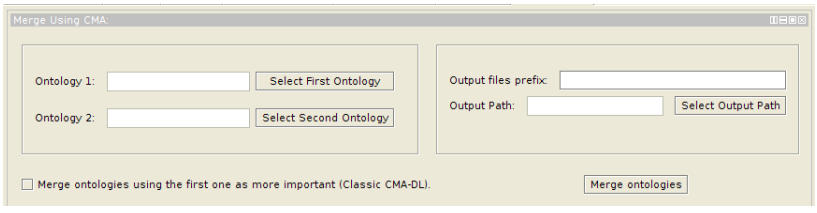


Figure 1. Protege Merging Plugin

[5] `http://hermit-reasoner.com/`
[6] `http://owlapi.sourceforge.net`

4. Conclusions and Future Work

In this paper, we presented a brief overview of theoretical and pragmatic research in the ontology merging field. We showed that there is a big gap between the pragmatic and theoretical approaches for ontology merging. We believe that both sides would benefit from a higher degree of integration. Also, we believe that the pragmatic field is a little bit stagnant when it comes to dealing with inconsistent merging results. In that context, we have chosen the CMA-DL algorithm as a starting point for our research because it clearly solves inconsistencies and could be directly applicable to ontologies. After a few experiments, we have proposed a small modification for this algorithm to provide more power of choice to the users. Now, the user can choose which ontology ordering suits better his/her needs.

Currently, we are working on building a software library to manage inconsistencies. An initial version of it is available at `http://ccsl.ime.usp.br/en/onair/ontology-merging`. Also, we intend to implement a web version for this merging mechanism and integrate it to the OnAIR - Ontology Aided Information Retrieval system[7], which is an ontology-based search tool for multimedia bases. OnAIR has an ontology-based query expansion feature and the merging mechanism would help the experts to build better ontologies, improving the retrieval results quality.

References

Falconer, S., Noy, N., and Storey, M. (2007). Ontology mapping-a user survey. In *Proceedings of the Workshop on Ontology Matching (OM2007) at ISWC/ASWC2007*, Busan, South Korea.

Gorogiannis, N. and Hunter, A. (2008). Merging first-order knowledge using dilation operators. In *Proceedings of the 5th international conference on Foundations of information and knowledge systems (FoIKS'08)*, pages 132–150, Berlin, Heidelberg. Springer-Verlag.

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

Konieczny, S. and Pérez, R. P. (1999). Merging with integrity constraints. In *Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'99)*, pages 233–244.

Meyer, T., Lee, K., and Booth, R. (2005). Knowledge integration for description logics. In *Proceedings of the National Conference on Artificial Intelligence AAAI'05*, volume 20, pages 645–650. AAAI Press.

Noy, N. and Musen, M. (2000). Prompt: algorithm and tool for automated ontology merging and alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00).*, pages 450–455.

Qi, G., Liu, W., and Bell, D. (2006). A revision-based approach to handling inconsistency in description logics. *Journal of Artificial Intelligence Review*, 26(1):115–128.

Revesz, P. Z. (1995). On the semantics of arbitration. *International Journal of Algebra and Computation*, 7:133–160.

van Harmelen, F., Haase, P., , Huang, Z., Stuckenschmidt, H., and Sure, Y. (2005). A framework for handling inconsistency in changing ontologies. *The Semantic Web–ISWC 2005*, pages 353–367.

[7] `http://ccsl.ime.usp.br/onair`

Uma ontologia de *engine* de jogos educativos para crianças com necessidades visuais: fase de preparação.

Romário P. Rodrigues[1], Gabriela R. P. R. Pinto[2], Cláudia P. P. Sena[3], Expedito C. Lopes[4], Teresinha F. Burnham[5]

[1 2 3]Área de Informática – Departamento de Exatas - Universidade Estadual de Feira de Santana (UEFS) – Feira de Santana – BA – Brasil

[4] Departamento de Informática - Universidade Católica de Salvador (UCSal) Salvador – BA - Brasil

[5] Faculdade de Educação (FACED) - Universidade Federal da Bahia (UFBA) Salvador – BA - Brasil

{romarioecomp, gabrielarprp, caupinto, ditoexpe}@gmail.com, tfroesb@ufba.br

Abstract. *This paper presents preliminary results of a end of course work from a computer engineering's student (EComp), which aims to develop an ontology engine for educational games for children with visual needs. In addition to the preliminary results, presents some methodological resources that are being used for developing such an ontology.*

Resumo. *Este artigo apresenta resultados preliminares do Trabalho de Conclusão de Curso (TCC) de um estudante do curso de Engenharia de Computação (EComp), que objetiva elaborar uma ontologia de engine de jogos educativos para crianças com necessidades visuais. Além dos resultados preliminares, apresenta os recursos metodológicos que estão sendo utilizados para o desenvolvimento da referida ontologia.*

1. Introdução

A condição de visão de pessoas com necessidades visuais pode ser categorizada como: *cegueira* e *baixa visão* [Glat 2009], também chamada de visão subnormal [Conde 2010]. Nos dois casos, tais necessidades afetam o desenvolvimento eficiente da visão para a execução de tarefas que exigem tal sentido. Naqueles que apresentam visão subnormal, o resíduo visual, em diferentes proporções, permite a essas pessoas o desenvolvimento de tarefas educacionais. Aqueles diagnosticados com cegueira conseguem, no máximo, perceber a luz [Gasparetto e Nobre 2007], dependendo, portanto, de outros recursos que os auxiliem em tais tarefas. Dessa maneira, o processo educacional envolvendo cegos requer atividades e recursos próprios, aplicação cuidadosa da mente e concentração [Rabello 2007]. No campo da Informática, diversas tecnologias assistivas (i.e *softwares,* máquinas) vêm sendo criadas com o intuito de colaborar no processo de formação e inclusão dos portadores de necessidades visuais nos diversos setores da sociedade.

O Ministério da Ciência e Tecnologia (MCT), através da Pesquisa Nacional de Tecnologia Assistiva (PNTA), tem acompanhado a produção de tecnologias assistivas e fornecido, a cada ano, dados estatísticos sobre a sua produção, necessidades, uso e investimento no Brasil [MCT 2011]. Conforme as informações divulgadas no portal MCT (2011), o uso de tecnologias para pessoas com necessidades visuais tem ganhado importância nas mais diversas áreas do conhecimento humano. Um exemplo de tecnologia indicada para promoção da interatividade e aprendizagem para cegos são os jogos eletrônicos (*e.g aShooter; BlastChamber; Bobby'sRevenge; Bop It Ultimate; CrazyDarts; CrazyTennis; Deekout DescentintoMadness*).

O motor de jogos eletrônicos, também conhecido pela palavra inglesa *engine*, é o *software* responsável por carregar os arquivos de arte para a memória, realizar os desenhos, tocar sons, etc. Ele oferece componentes reutilizáveis que podem ser manipulados para carregar, exibir e animar modelos, e que podem ser utilizados para a produção de outros jogos com características básicas similares [Sanches 2010].

Todavia, percebe-se a escassez de *engines* para jogos eletrônicos educativos para crianças cegas. Tomando isto como uma demanda de pesquisa, e motivado pela possibilidade de desenvolver habilidades e competências técnicas fundamentais para o seu desenvolvimento profissional e poder contribuir com a inclusão social, Rodrigues (2011) está objetivando compreender os benefícios advindos da elaboração de uma modelagem de um *engine* de jogos eletrônicos educativos para crianças cegas.

Para a consecução do objetivo de pesquisa supracitado, foram realizadas algumas ações tais como o levantamento, exploração e análise de jogos eletrônicos educativos disponíveis para crianças cegas; a identificação de requisitos funcionais e não funcionais que são comuns aos jogos explorados; e a representação do domínio *engine* de jogos eletrônicos educativos para crianças cegas, a partir da elaboração de uma ontologia.

Ontologia, conforme explicam Gruber (1993) e Guarino (1998), é uma especificação explícita e formal de uma conceitualização compartilhada. E Breitman (2005) explica essa definição, comentando que a "conceitualização" representa um modelo abstrato de algum fenômeno que identifica os conceitos relevantes para o mesmo; "especificação explícita" significa que os elementos e suas restrições estão claramente definidos; "especificação formal" significa que a ontologia deve ser passível de processamento automático; e o adjetivo "compartilhada" reflete a noção de que uma ontologia captura conhecimento consensual, aceito por um grupo de pessoas. Segundo Breitman (2005), as ontologias têm sido adotadas por diversas comunidades formadas por profissionais de diversas áreas de Engenharia de Computação, como Inteligência Artificial, Representação do Conhecimento, Processamento de Linguagem Natural, Web Semântica, Engenharia de Software, entre outras.

Guarino (1998) afirma que é possível classificar as ontologias quanto à generalidade em ontologias de nível superior, que descrevem conceitos muito genéricos (e.g. espaço, tempo e eventos); ontologias de domínio, que descrevem o vocabulário relativo a um domínio específico através da especialização de conceitos presentes na ontologia de alto nível; ontologias de tarefas, que descrevem o vocabulário relativo a uma tarefa genérica ou atividade através da especialização de conceitos presentes na ontologia de alto nível; e ontologias de aplicação, que são ontologias mais específicas.

Este artigo objetiva apresentar alguns passos que estão sendo dados no intuito de elaborar a ontologia de domínio *engine* de jogos eletrônicos educativos para crianças cegas. Aqui, o termo "cego" é utilizado tanto para se referir às pessoas que apresentam baixa visão, quanto aos portadores de cegueira total. O artigo está articulado da seguinte forma: a Seção 2 apresenta algumas escolhas metodológicas que foram realizadas no intuito de desenvolver a ontologia. Os resultados preliminares, obtidos a partir da execução da primeira fase da metodologia de desenvolvimento de ontologias que foi adotada, são apresentados na Seção 3; e, finalmente, na Seção 4, as considerações finais são elencadas e os trabalhos futuros apontados.

2. Escolhas metodológicas para o desenvolvimento da ontologia

Para o desenvolvimento da ontologia de *engine* de jogos eletrônicos educativos para pessoas cegas, utilizaram-se a *Ontology Web Language* (OWL) e o editor *Protegé*.

A *Ontology Web Language* (OWL) apresenta como principais elementos para a construção de uma ontologia: (1) *Namespaces*: são declarações que permitem que os identificadores que estão presentes na ontologia sejam interpretados sem ambigüidades; (2) Classes: representa um conjunto ou coleção de indivíduos que servem para descrever conceitos de um domínio; (3) Indivíduos: são membros das classes e que podem se relacionar a outros indivíduos através de propriedades; (4) Propriedades: que são relacionamentos binários e sevem para descrever fatos em geral, a todos os membros de uma classe, ou a um indivíduo dessa classe; e (5) Restrições: são utilizadas para definir alguns limites para indivíduos que pertencem a uma classe [Breitman 2005].

O *Protege* é um software livre, de código aberto, que fornece, para uma comunidade de usuários em crescimento, um conjunto de ferramentas para a construção de modelos de domínio e bases de conhecimento baseadas em ontologias. Em seu núcleo, implementa um rico conjunto de estruturas de modelos de conhecimento e suporta a criação, visualização e manipulação de ontologias em diversos formatos e representação. Pode ser personalizado para fornecer apoio amigável para elaboração de modelos de conhecimento. Além disso, pode ser estendido por meio de um *plug-in* e uma arquitetura baseada em *Java Application Programming Interface* (API) para elaboração de ferramentas baseadas em conhecimento e aplicações [Protege 2011].

Além da adoção da linguagem OWL e do Protege, foi escolhida a Metodologia de Desenvolvimento de Ontologias (MDO) para auxiliar Rodrigues (2011) quanto aos procedimentos metodológicos necessários para a elaboração da ontologia. A MDO foi proposta por Pinto et *al*. (2005) e está articulada em quatro fases: (1) Preparação, (2) Formalização, (3) Implementação e (4) Análise de Qualidade: A Fase de Preparação está dividida em três processos (Análise do Domínio, Levantamento de Conceitos e Identificação de Relacionamentos e funções) os quais objetivam ter um modelo conceitual do domínio que está sendo analisado. A Fase de Formalização está dividida em dois processos (Definição de Axiomas e Mapeamento Estrutural), que definem a lógica proposicional e esquematizam o modelo conceitual, definido na Fase de Preparação. A Fase de Implementação executa a ontologia. Esta tem um processo (Prototipação), o qual possibilita uma possível modificação em sua ontologia. Na Fase de Análise de Qualidade, os processos de Validação e Verificação e Usabilidade são utilizados para verificação de falhas e da aplicabilidade da ontologia construída.

Esta pesquisa encontra-se na Fase de Preparação da MDO. Alguns jogos eletrônicos foram selecionados, explorados e analisados, de modo que o estudante pudesse compreender o domínio de conhecimento e levantar conceitos candidatos a integrar a ontologia. Em seguida, os conceitos candidatos foram relacionados. A representação gráfica gerada a partir de *OWLViz*, do *Protegé*, será apresentada na próxima seção.

3. Resultados Preliminares

Atualmente, a ontologia elaborada contém 120 conceitos, obtidos por meio da análise de 20 jogos destinados a pessoas cegas. Para analisar os conceitos existentes nos jogos, tomou-se como referência o trabalho de Zagal (2010a, 2010b), que, em seu projeto intitulado *Game Ontology Project* (GOP), objetiva representar as características dos jogos em conceitos.

A fim de apresentar as características básicas da modelagem do *engine* proposto, apresentam-se, neste artigo, alguns dos conceitos obtidos. Eles encontram-se relacionados taxonomicamente a partir do conceito raiz *OntoEngineGameBlind,* são eles: *Interface, Rules, EntityManipulation, Goals*, conforme pode ser observado na Figura 1.

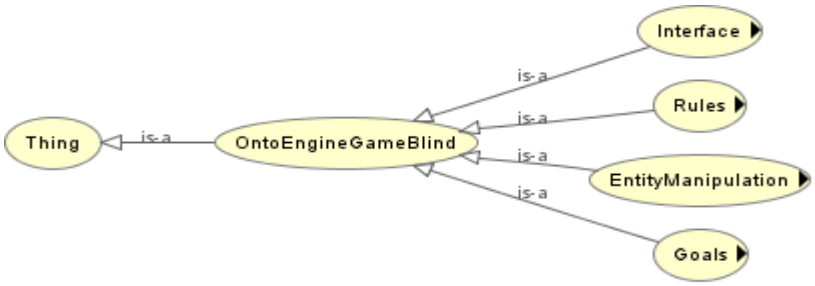


Figura 1. Definições básicas da Modelagem.

Fonte: [Rodrigues 2011]

Interface é um conceito onde se encontram a apresentação do jogo, as entradas de dispositivos e seus métodos. Divide-se basicamente em Entradas (*Input*) e Apresentação. Os *Inputs* utilizam de componentes que interagem com o usuário através de dispositivos físicos, e o *Presentation* dispõe na tela os resultados e as possibilidades existentes para que a pessoa utilize o jogo. A Figura 2 expõe a Interface e seus subníveis.

Rules (Regras) definem e estabelecem as condições de um jogo a serem seguidas. As regras são divididas em Regras do Jogo, Regras do Mundo do Jogo e as Regras de Sinergia, estas últimas utilizadas para a comunicação das duas anteriores. A Figura 3 mostra o conceito *Rules* e seus subníveis.

Entity Manipulation, ou Manipulação de Entidades, realiza as ações principais de um jogador. Estas ações podem ser: a customização de jogadores e cenários, a movimentação do jogador, a composição de ações dentro de um jogo, a criação, remoção e seleção de componentes e jogadores, além da possibilidade de acontecer

colisões em um jogo e a manipulação do tempo. A Figura 4 mostra os conceitos relacionados à Manipulação de Entidades.

Goals são os objetivos de um jogo, as metas a serem alcançadas. Os objetivos podem ser os do jogo (*GameGoals*), que tem suas próprias limitações, os limites para cada objetivo(*GoalMetrics*), e os objetivos do usuário (jogador) do jogo (*AgentGoals*). A Figura 5 exibe o conceito *Goals* e seus subníveis correspondentes.

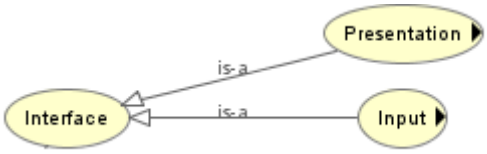


Figura 2. *Interface* e seus subníveis.
Fonte: [Rodrigues 2011].

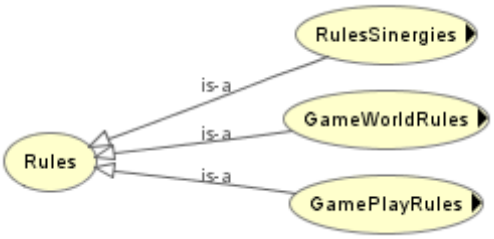


Figura 3. *Rules* e seus subníveis.
Fonte: [Rodrigues 2011].

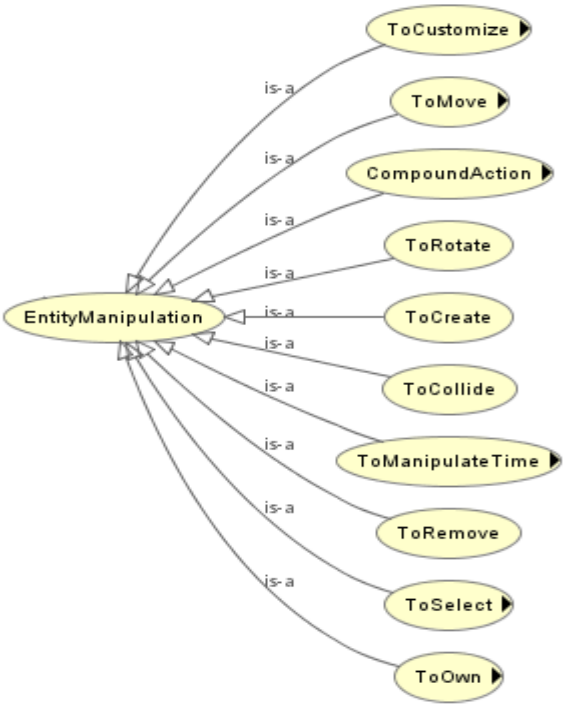


Figura 4. *Entity Manipulation* e seus subníveis.

Fonte: [Rodrigues 2011]

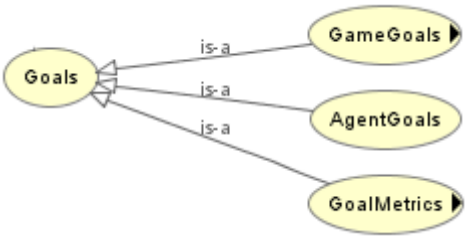


Figura 5. *Goals* e seus subníveis.
Fonte: [Rodrigues 2011].

4. Considerações Finais

Este trabalho objetivou apresentar alguns resultados preliminares tendo como base um Trabalho de Conclusão de Curso (TCC) [Rodrigues 2011]. Com o intuito de contribuir com a inclusão social e desenvolver habilidades técnicas fundamentais para a sua formação profissional, está sendo desenvolvida a ontologia *engine* de jogos eletrônicos educativos para crianças cegas. Planeja-se dar continuidade ao trabalho

através da execução das fases de Formalização, Implementação e Análise de Qualidade da MDO, com o intuito de finalizar o desenvolvimento da ontologia *engine* de jogos educativos para crianças cegas, e disponibilizá-la para uso.

Referências

Breitman, K. K. (2005) "Web Semântica: A internet do futuro". Rio de Janeiro: LTC.

Conde, A. J. M. (2010) "Definindo a Cegueira e a Visão Subnormal". IBC (Instituto Benjamin Constant), http://www.ibc.gov.br/?itemid=94#more>. Dezembro.

Gasparetto, M. E. R. F.; Nobre, M. I. R. de S. (2007) "Avaliação do Funcionamento da Visão Residual: Educação e Reabilitação". In: MASINI, E. F. S. (organizadora): A Pessoa com Deficiência Visual: um Livro para Educadores. 1ª edição, São Paulo: Vetor.

Glat, R. (2009). "Educação Inclusiva: Cultura e Cotidiano Escolar". Rio de Janeiro: Letras, Ed.Viveiros de Castro Ed. Ltda.

Gruber, Tom (2010). "*What is an Ontology*". http://www-ksl.stanford.edu/kst/what-is-an-ontology.html. Maio.

Guarino, N. (1998), "Formal Ontology and Information Systems". http://www.loa-cnr.it/Papers/FOIS98.pdf. Maio.

Pinto, G. R. P. R. et *al.* "Definição de uma ontologia para os canais preferenciais de difusão do conhecimento técnico-científico: Fase de preparação". In: Fróes Burnham T. et *al*. (Org.). Mosaico: Difusão do Conhecimento na Sociedade da Aprendizagem. Salvador; Feira de Santana: EDUFBA; UEFS, b, v. 1, p. 137-147.

Protege (2011). "What is protégé?" http://protege.stanford.edu/overview/. Setembro.

Rabello, R. S. (2007). "O Teatro na Educação do Deficiente Visual e a Teoria da Peça Didática de Brecht. Revista da FAEEBA: Educação e Contemporaneidade", Salvador, v. 16, n. 27, p. 163-164.

Rodrigues, R. P. (2011). Uma ontologia de *engine* de jogos educativos para crianças com necessidades visuais. Trabalho de Conclusão de Curso. Curso de Engenharia de Computação, Universidade Estadual de Feira de Santana (UEFS), Feira de Santana, Bahia.

Sanches, B. C (2010). Os softwares de um jogo. http://www.pontov.com.br/site/index.php?option=com_content&view=article&id=108:os-softwares-de-um-jogo&catid=51:programacao&Itemid=65. Julho.

Zagal, J. *et al*. (2010a) "The Game Ontology Project: Supporting Learning While Contributing Authentically to Game Studies". http://portal.acm.org/ft_gateway.cfm?id=1599933&type=pdf. Julho.

Zagal, J. *et al.* (2010b) "Towards an Ontological Language for Game Analysis". http://users.soe.ucsc.edu/~michaelm/publications/zagal-digra2005.pdf. Agosto.

SUPORTE DE ONTOLOGIAS APLICADAS À MINERAÇÃO DE DADOS POR REGRAS DE ASSOCIAÇÃO

Eduardo de Mattos Pinto Coelho[1], Marcello Peixoto Bax[2], Wagner Meira Jr.[3]

[1]Prefeitura Municipal de Belo Horizonte

[2]Escola de Ciência da Informação. Universidade Federal de Minas Gerais (UFMG)

[3]Departamento de Ciência da Computação, (UFMG)

(ematos@pbh.gov.br, bax@eci.ufmg.br, Meira@dcc.ufmg.br)

Abstract.

Data Mining (DM) for association rules tends to generate an unmanageable number of rules affecting the scope of its application. To solve this problem we propose the use of ontologies in the stages of pre and post-processing tasks to support the MD. In addition, the article points out that human organizations require the notions of possibility, subjectivity and interpretation, contrasting with the notions of necessity, objectivity and explanation, useful in fields of natural sciences. These requirements demand new perspective on ontologies and DM, often sheltered by soft computing.

Resumo. *Mineração de dados(MD) por regras de associação tende a gerar um número intratável de regras prejudicando a abrangência de sua aplicação. Para solucionar esse problema propõe-se o uso de ontologias nas etapas de pré e pós-processamento no suporte às tarefas de MD. Além disso, o artigo ressalta que organizações humanas exigem as noções de possibilidade, subjetividade e interpretação, contrastantes com as noções de necessidade, objetividade e explicação, úteis em domínios de ciências naturais. Tais exigências demandam novas perspectiva em ontologias e MD, normalmente abrigadas pela computação suave.*

1. Introdução

Ontologias e mineração de dados (MD) são áreas, em geral, construídas com base nas ciências naturais (CN). Lidam com as noções de necessidade, objetividade e explicação, em ciências físicas, químicas e biológicas. Entretanto, em organizações, fenômenos humanos prevalecem, demandando metodologias de ciências humanas (CH) que, por sua vez, exigem as noções de possibilidade, subjetividade e interpretação. Tais noções já são abrigadas por tecnologias humano-cêntricas na perspectiva da computação suave. Assim, com base em sistemas *difusos*, propõe-se utilizar ontologias no suporte à atividade de MD nas fases de pré e pós-processamento. Agrega-se então valor conceitual, associado ao conhecimento de domínio, e propicia-se suporte semântico no processamento dos dados. O analista é auxiliado na explicação e interpretação das regras obtidas da MD. Inicialmente o artigo introduz noções de MD. A partir da análise das tarefas de prescrição, predição, descrição, explicação e interpretação, conforme abordados em Domingues (2004), são discutidos métodos que devem ser incorporados a ontologias em contextos humano-sociais, típicos de questões surgidas em economia, administração, contabilidade, direito, etc.

2. Mineração de Dados

A MD é uma das etapas da Descoberta de Conhecimento em Bases de Dados. É um processo de identificação de padrões novos, potencialmente úteis e compreensíveis, em um conjunto de dados. Esses padrões podem se constituir, dentre outros, em regras de associação, ou seqüências temporais que permitam revelar relacionamentos não aleatórios entre atributos de variáveis. Em especial, o volume excessivo de dados justifica a MD para a descoberta de padrões que possam revelar informações úteis à tomada de decisão.

A MD lidará com questões de ordem computacional e de usabilidade. Ter-se-á de lidar com a confiabilidade dos dados, a geração de resultados excessivos, autonomia e conhecimento do negócio para a análise dos resultados, desempenho e produtividade, facilidade de uso e utilidade.

Há uma variedade de paradigmas de mineração de dados. Objetivando nossa análise, consideramos mais interessante a taxonomia que realça a existência de modelos supervisionados e não-supervisionados.

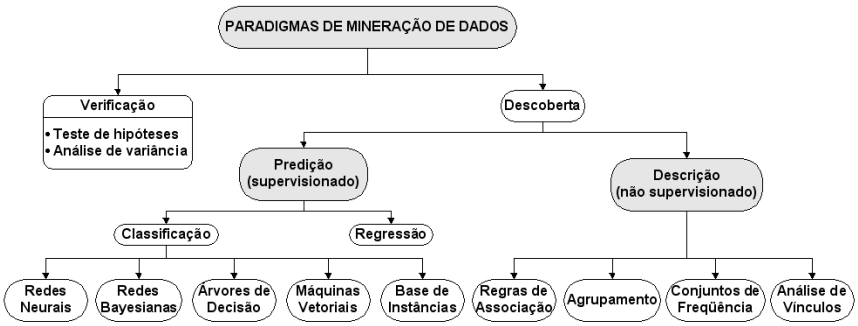


Figura 1. Taxonomia de paradigmas de mineração de dados (Maimon e Rokach, 2005, p. 7).

Em **modelos** supervisionados, como o próprio nome já diz, o processo de detecção de padrões é supervisionado. As classes à qual cada padrão possa ser pertinente é pré-definida. São **modelos prescritivos**. Além disso, são preditivos no sentido de serem mais adequados a revelar tendências.

Já em modelos não-supervisionados, não se conhecem a priori classes às quais os padrões possam ser pertinentes. São **modelos de ênfase descritiva**, no sentido de serem mais adequados a apresentarem descrições, relevando características dos dados minerados.

3. Ontologias no suporte à MD

A despeito dos algoritmos de MD atenderem critérios de desempenho na criação de novos dados, freqüentemente, eles são insuficientes para garantir seu uso prático. Conhecer o domínio é crucial. Vem-se desenvolvendo um paradigma MD guiado pelo conhecimento, Cao (2010). Em reforço a esse paradigma, ontologias são utilizadas. Ontologias no pós-processamento de regras de associação aparece em Marinica, Guillet e Briand (2009) e Marinica e Guillet (2010).

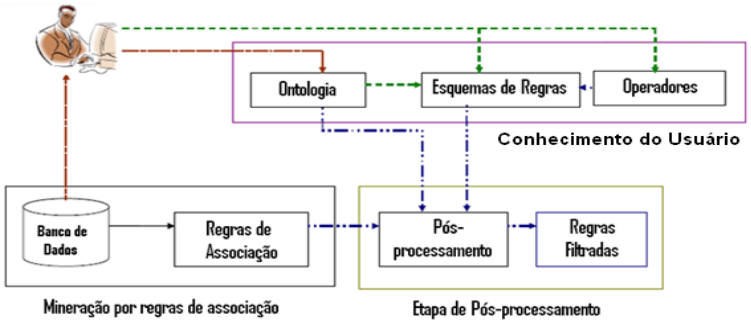


Figura 2. MD com suporte de ontologias, esquemas de regras e operadores.

4. Peculiaridades das ciências humanas

A análise científica de fenômenos humano-sociais deve articular dialeticamente os níveis descritivo, explicativo e interpretativo. Como apresentado na Figura 1, a MD subdivide-se em dois ramos: mineração supervisionada e não-supervisionada. A supervisão está associada à atividade de previsão, a partir de conhecimentos explicitamente pré-estabelecidos (**prescrição**), descobrem-se fatos novos, mas de algo que já se conhecia. A não-supervisão é associada à atividade de **descrição**. Não há conhecimento prévio. Busca-se a descoberta de associações a serem descritas, conhecimento útil. Em mineração supervisionada, ontologias, podem ser vistas tanto para representar conhecimento previamente adquirido, necessário à previsão, quanto para descrever conhecimento útil em minerações preditivas posteriores. Na mineração não-supervisionada, ontologias suportam a análise dos padrões descritivos obtidos no processo de mineração, tarefas de **explicação** e **interpretação**. A ênfase na explicação é típica das CN, enquanto na tarefa de interpretação é típica das CH (Domingues, 2004, p. 103 a 135, em especial, p. 116 e p. 134-135).

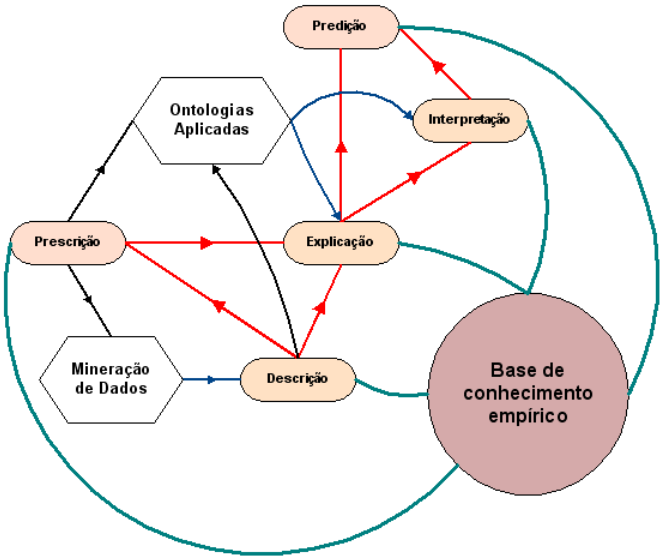


Figura 3. Suporte de ontologias e MD às tarefas que se articulam dinâmica e dialeticamente para a constituição e no uso da base de conhecimento empírico.

Com base nessas características da MD e de ontologias, a Figura 3 mostra como MD e ontologias aplicadas se inserem no auxílio às tarefas de predição, interpretação, explicação, descrição e prescrição. A seguir iremos esclarecer como a consideração dessas tarefas revela que os métodos de constituição de bases de conhecimento, em CH, possuem peculiaridades que os distinguem dos voltados às CN.

4.1. Tensão entre a objetividade inalienável *versus* a subjetividade inexorável

Ao se buscar referências para o desenvolvimento ontologias específicas para domínios humano-sociais, de imediato, depara-se com questões de fundamentação. O método em CH é ainda questão aberta. Seus padrões de cientificidade encontram-se em busca de uma consolidação a partir de distintas vias. Seus fenômenos resistem às exigências da objetividade. No cerne dos dilemas do método, na busca inalienável da objetividade, encontra-se a dificuldade de objetivação da significação dos fatos sociais e de lidar com a subjetividade inerente aos fenômenos humano-sociais. Se a objetividade é a meta inalienável, a subjetividade é elemento inexorável que se imiscui nos fatos, nos fenômenos, e na práxis das CH.

4.2. Descrição

Um mesmo fato, ou fenômeno, pode ser descrito de diferentes maneiras e a descrição nunca será completa. Ela depende da granularidade observada, e da intencionalidade do observador. A intencionalidade condiciona a granularidade. Mesma coisa nas CN. Peculiar das CH é que, além dos aspectos objetivos dos comportamentos humanos, a descrição deverá debruçar-se sobre um conjunto de elementos subjetivos

(intenções, sentimentos, consciência, valores e fins visados pelos agentes humanos) (Domingues, 2004, p. 107). As CN debruçam-se sobre causas, descrição e explicação causal dos fatos. As CH estendem esse leque para buscar a descrição e explicação das motivações, razões e crenças. A descrição dever-se-á, necessariamente, lançar seus elementos à tarefa interpretativa que, em CH, raramente já estará encapsulada pela tarefa explicativa.

4.3. Explicação

Com base nos elementos descritivos há distintas formas de explicação possíveis: genéticas, estruturais, funcionais, finais e, dentre outras, causais. **A explicação causal é considerada a forma de explicação por excelência** (Domingues, 2004, p. 116). Entretanto, a questão da causalidade, assim como a questão da vaguidade, é assunto de intensos debates. A causalidade aproxima-se da vaguidade. Assim como ela, a noção de causalidade tem sido útil para fins de análise e, como a vaguidade, ao invés de ser abandonada, ressurge revigorada nas últimas décadas. Uma mesma coisa pode ser causa de efeitos contrários; pode-se identificar uma causalidade recíproca ou circular; pode ocorrer dependência mútua, ação, ou influência de causa e efeito e, ainda, causalidade freqüentemente é confundida com condicional lógica. Conclui Domingues (2004, p. 119) que o importante é a análise causal depender da consideração de um contexto mais amplo, que se decide em outro nível de análise: a interpretação.

4.4. Interpretação (compreensão)

Muitos pretendem que a interpretação já se decide no nível da explicação, e não é senão um de seus aspectos. Entretanto, para Domingues, é o caso de distinguir uma da outra, considerando que a explicação incide sobre os fatos, ou coisas. Já a interpretação envolve a significação, o sentido deles. Portanto, a interpretação irá introduzir as unidades significativas de análise, como as hipóteses, os modelos (tipos ideais), as postulações de sentido, e assim por diante (Domingues, p. 119-120). Em CH, a candidata de ter a primazia no método e de conduzir a análise é a tarefa interpretativa.

4.5. Inter-subjetividade

A subjetividade é crucial na análise do conhecimento em domínios de CH. Anscombe considera não só a subjetividade de um indivíduo, mas a de coletividades inteiras. Desafio maior, é a inter-subjetividade; comunicação das consciências individuais. Na construção de ontologias, a inter-subjetividade é tratada. Ontologias expressam consenso.

Outro lado da questão é o uso de ontologias na etapa de decifração do sentido intersubjetivo na tarefa da interpretação. A verdade a ser considerada advirá da convicção de verdade recolhida pela análise na comunidade. Em CI e CC, a demanda pela inter-subjetividade é eventual. Mas tal demanda é sentida nas comunidades de inteligência onde a inter-subjetividade é regra. Distintos analistas são confrontados numa análise conjunta que traduza a convergência das análises individuais.

5. Quais fundamentos em ontologias demandam as ciências humanas?

Peculiaridades das CH em relação as CN são: primazia da interpretação sobre a explicação; importância e necessidade de subjetividade em conjunto com a objetividade; flexibilização da noção de necessidade (lei) para a noção de possibilidade. Ao buscar agregar conhecimento de domínio, via ontologias, no suporte à MD em contextos organizacionais, impõe-se a nós duas questões: (1) As ontologias fundamentais, ou de alto nível, referências na construção de novas ontologias, lidam com peculiaridades

atinentes ao domínio de CH? (2) Como ontologias poderiam incorporar as noções de subjetividade e possibilidade, suportando a tarefa de interpretação? Não examinaremos essas questões, mas teceremos a seguir alguns comentários, indicações e avaliações iniciais que poderão nortear avaliações mais aprofundadas no futuro.

5.1. Ontologias Fundamentais[1].

DOLCE, a SUMO e a BFO explicitam suas orientações filosóficas. Consideram universais, particulares, ou tropos, seu compromisso ontológico, entidades no tempo e no espaço, e como se relacionam. Oberle et. al (2007) avalia as citadas acima, com base em quatro pares de escolhas ontológicas. Descritiva, ou revisionária (prescritiva); multiplicativa, ou reducionista; atualista, ou possibilista; e endurante, ou perdurante. Considerando o subjetivismo e a noção de possibilidade inerentes às CH, avalia-se que escolher o descritivismo, o multiplicativismo e o possibilismo permite melhor tratar dessas peculiaridades. O descritivismo, ao considerar o que a realidade é, e não como deveria ser, está apto a tratar da linguagem natural, do senso comum, da vaguidade, de objetos não físicos, da diferenciação entre objetos e processos etc. O multiplicativismo por considerar a diferenciação de entidades a partir do não compartilhamento de propriedades fundamentais, também está mais próximo ao senso comum. O possibilismo, por levar à modalidade que faz uso da possibilidade.

Comenta-se brevemente as três ontologias. Ver detalhes em Sreejith (2008). A DOLCE adota uma metafísica descritiva, baseada em Strawson[2], em oposição à metafísica prescritiva, e considera aspectos lingüísticos e de engenharia cognitiva. Procura incorporar em sua estrutura elementos que permitam lidar com artefatos cognitivos, marcas culturais e convenções sociais (um tipo de metafísica cognitiva). Inspira-se na noção de *deep background* desenvolvida por Searle (2002, *Intencionality*). A SUMO vai na mesma direção, classificando atos intencionais[3], orientações, interações sociais (Sreejith, 2008, p. 66). Oberle et. al (2007), em solução híbrida (SWIntO), integra a DOLCE e a SUMO. Acomoda resistências à DOLCE, abstrata demais, e à SUMO de axiomatização árdua. A DOLCE teria uma proposta mais abrangente, já a SUMO teria uma taxonomia mais rica. Por fim a BFO distingue entre dois tipos de entidades: substanciais ou continuantes, e processuais, ou ocorrentes. Smith considera a BFO um subconjunto de DOLCE, adequada ao tratamento de instâncias, tipos e relações. Mais adequada que a SUMO, que, segundo esse autor, não apresenta um tratamento claro de relações entre instâncias *versus* relações entre tipos (Barry Smith, *Upper Level Ontologies*). Em CH, a DOLCE e a SUMO já são ao menos parcialmente aptas a lidar com as peculiaridades apresentadas anteriormente, enquanto que a BFO vai na contramão das considerações ontológicas envolvendo aspectos do senso comum, da linguagem natural, da consideração de atos intencionais, vaguidade, causação mental etc. Consideramos que, a despeito dos riscos, é inevitável, para o bem da própria objetividade científica, buscar lidar e tratar de questões que se nos impõem por sua força e penetrabilidade em amplos domínios, envolvendo aspectos cognitivos, tais como a vaguidade, a subjetividade, a intenção e a causação mental. A BFO não pretende alcançar domínios das CH. Ver debate recente (Merril, 2010a e 2010b *versus* Smith e Ceusters, 2010).

---

[1] *Foundational Ontologies*, por vezes traduzidas como ontologias "fundacionais", um neologismo. Também são chamadas de ontologias de alto nível (*upper ontologies*).

[2] STRAWSON, P.F.. "Individuals. *An Essay in Descriptive Metaphysics*. Routledge, 1959.

[3] http://virtual.cvut.cz/ksmsaWeb/browser/print/3%23IntentionalProcess e http://swserver.cs.vu.nl/partitioning/SUMO/

5.2 A emergência da computação suave, granular e humano-cêntrica

A insuficiência das soluções focadas em detalhes técnicos como eficiência de algoritmos, vem levando à consideração da usabilidade, conhecimento do domínio e capacidade de interpretação de resultados. A computação suave e granular se associa à uma abordagem humano-cêntrica, flexível e com foco na interação do homem com a máquina. Considerando-se essas três abordagens convergentes – computação suave, granular e humano-cêntrica – destacam-se as soluções em sistemas difusos, redes neurais e algoritmos genéticos. Mitra (2002) dá uma visão da literatura disponível sobre MD. Sistemas difusos preocupam-se com a natureza amigável ao usuário (Pedrycz e Gomide, 2007, p. xvii). Sistemas difusos podem ser integrados com outras ferramentas de computação suave levando à geração de sistemas mais poderosos, com aplicações em reconhecimento de padrões, processamento de imagens, e inteligência de máquina (Mitra e Pal, 2005). Nesse contexto, confirmando a emergência dessas abordagens, vem surgindo estudos que aplicam tecnologias de sistemas difusos a MD e ontologias.

6. Conclusão

As ferramentas computacionais de MD, por si só, não garantem o sucesso em ambientes organizacionais. Esses ambientes sócio-humanos possuem peculiaridades e exigem abordagens distintas daquelas associadas às CN. Ao invés do caráter de lei associada à noção de necessidade, há a noção de possibilidade; ao invés da ênfase na explicação, há a ênfase na interpretação, e ao invés do expurgo da subjetividade, há a sua concomitante consideração, não em contraposição, mas em cooperação com a objetividade científica. O surgimento de novas abordagens apresentadas traz novas metodologias que permitem tratar as peculiaridades dos domínios sócio-humanos.

7. Bibliografia

CAO, Longbing. Domain-Driven Data Mining: Challenges and Prospects. Transaction on Knowledge And Data Engineering, Vo. 22, n° 6, June, 2010, pp. 755-769.

DOMINGUES, Ivan. *Epistemologia das Ciências Humanas. Tomo I: Positivismo e Hermenêutica – Durkheim e Weber*. Edições Loyola, 2004, 671 p.

MAIMON, Oded e ROKACH, Lior. "Introduction To Knowledge Discovery in Databases". Chapter 1, In: *The Data Mining and Knwoledge Discovery Handbook*, Springer, 2005, p. 1-17,

MARINICA, Claudia e GUILLET, Fabrice. "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 6, Feb. 2010, pp. 784-797.

MARINICA, C., GUILLET, F., & BRIAND, H. Post-Processing of Discovered Association Rules Using Ontologies. *Proc.2008 IEEE Intern Conf. on Data Mining Workshops*, 2009, pp.126–133.

MERRILL, G.H. "Ontological realims: Methodology or misdirection?" *Applied Ontology*, V. 5, 2010, pp. 79-108.

MERRILL, G.H. "Realism and reference ontologies: Considerations, reflections and problems". *Applied Ontology*, Vol 5, 2010, pp. 189-221.

MITRA, S. e PAL, S. K. Fuzzy sets in pattern recognition and machine intelligence. Fuzzy Sets And Systems, 156, 2005, pp. 381-386.

OBERLE, Daniel et al. DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology). Junho, 2007

PEDRYCZ, Witold e GOMIDE, Fernando. Fuzzy Systems Engineering – Toward Human-Centric Computing. IEEE Press, John Wiley & Sons, 2007.

SEARLE, John R. Intencionalidade. Martins Fontes, 2002, 390 p.

SMITH, Barry e CEUSTERS, Werner. "Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology*, Vol 5, 2010, pp. 139-188.

SREEJITH, A. A Project Report on Neo-Vaisesika Formal Ontology. Department Of Computer Science, Cochin University of Science & Tecnology, Kochi, 2008.

A representational framework for visual knowledge

Alexandre Lorenzatti[1], Carlos E. Santin[2], Oscar Paesi da Silva[1], Mara Abel[1]

[1]Department of Computer Science – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]ENDEEPER
Porto Alegre – RS – Brazil

{alorenzatti,marabel,oscar.paesi}@inf.ufrgs.br, carlos.santin@endeeper.com

Abstract. *Visual knowledge refers to the set of conceptualizations owned and used by imagistic-domain experts on problem-solving tasks. Because of its visual nature, there is a lack of constructs for representing this kind of knowledge using ontologies. This paper presents hybrid meta-constructs proposed to formalize and represent visual knowledge using ontologies. Beyond that, this paper presents a visual knowledge based system which uses a domain ontology and the meta-constructs.*

1. Introduction

The aim of our research is to create appropriate models to formalize knowledge in order to support reasoning on knowledge intensive domains. Lately, our research is focused on the creation of visual knowledge models applied to imagistic domains.

Imagistic Domains are the ones where the domain expert starts the problem-solving process with a visual pattern matching over visual information input, which will further support the more abstract processes of inference. Image-based diagnosis in Medicine, and visual analysis of petroleum-reservoir rocks are common tasks executed by domain experts and are highly based on visual-information input.

Visual knowledge refers to the set of conceptualizations owned and used by individuals to recognize the relevant features in the domain and start the inference process. This kind of knowledge is built through the experience and differs from the propositional knowledge in the sense that the expert is not able to either express it verbally or in a sentential manner [Abel et al. 2005]. Since ontologies formalize knowledge by making use of propositional vocabulary, it is showing itself a challenge to formalize and represent visual knowledge using ontologies.

Visual knowledge and image are disjointed concepts. The Ullmann triangle [Ullmann 1979] describes the relation among an *Object* in the reality, a *Concept* in a conceptualization and a *Symbol* in a language (its extended version is presented on Figure 1, firstly published in [Lorenzatti et al. 2011]). An *Object* is supposed to be a real or concrete object where its existence can only be referred by the perception process of someone. *Concepts*, by their side, are abstractions that humans create over objects in order to deal with the external world. *Symbols* are the trial of individuals for representing concepts during the externalization process of communication in order to share a conceptualization among a community. A concept can have different representations and they

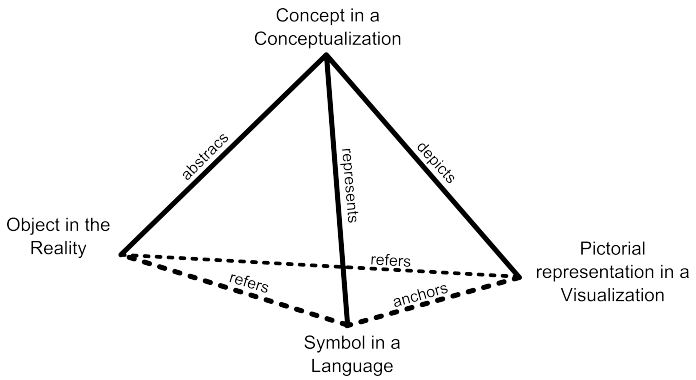


Figure 1. The pictorial representation vertex added in the extended version of the Ullmann triangle.

will vary according to the purpose of the chosen language. The image concept is normally referred as a pictorial representation of a concept.

When mentioning *visual knowledge* we are referring to the *conceptualization* vertex of the Ullmann triangle. A pictorial representation is an alternative representation that can be used in the externalization process of communication. Its choice depends on the requirements for externalizing visual concepts, like in spatial or location problems that require visual representations for sharing concepts. Thus, we add the new vertex "Pictorial representation in a Visualization" to the Ullmann triangle (in the baseline of Figure 1). Thus, the same concept can be simultaneously represented by a propositional symbol in a language while being represented by a pictorial symbol in a visualization and this correspondence between them is called *anchoring*.

The visual knowledge that we are modeling does not reside on the image content, but into the conceptualization of the domain experts. Thus, we do not seek to model the abstract visual patterns found in the image content, but our objective is to model the set of visual concepts residing in the experts' mind.

2. Conceptual Modeling

Creating knowledge based systems requires deep systematicity in the engineering process when using ontologies to represent knowledge [Guarino and Welty 2002]. Guarino and colleagues [Guarino and Welty 2004] have proposed a systematical and domain independent methodology applied to evaluate and validate the ontological choices taken when building up an ontology. The analysis is based on ontological notions coming from philosophy and are represented by formal meta-properties.

We apply the unified foundational ontology - UFO - of Guizzardi [Guizzardi 2005] with the goal of orienting the semantic negotiation of the concepts and the consensus achievement among the interacting agents (artificial or human) within a community [Gangemi et al. 2002]. UFO was built by meta-concepts, like kind, role, phase, and mixin, whose definitions were based on formal meta-properties. The meta-properties like identity, rigidity, and uniqueness impose constraints over the taxonomical relations which prevent the creation of inconsistent knowledge models.

The meta-property *identity* refers to the problem of identifying a single instance

based on its intrinsic characteristics that, make it unique [Guizzardi 2005]. The *identity criterion* concept involves the analysis of conditions and characteristics which, for example, allow the identification of a person along the time. The *rigidity* meta-property is related to the essentiality of an individual having a property in order to preserve its identity [Guizzardi 2005]. *Being human* is an essential property to all human beings, otherwise they will loose their identity. *Being hard* is not essential to all instances of hammer since, hammer toys should not be essentially hard. The meta-property *uniqueness* deals with the problem of identifying the parts and limits of objects [Guizzardi 2005]. The meta-property is used to analyze the composition of an object in order to identify if it is a whole or a composition of other objects, for example. Instances of the property *being a lake* have well defined boundaries, while instances from the property *being water* have not. Thus, the analysis of the meta-property unity prevents modeling the property *being water* being subsumed by the property *being a lake*, which sounds intuitively correct. In fact a lake is not the proper water but, it is constituted by water.

Rigid sortals are concepts where individuals have ontological rigidity. A tree is a rigid sortal, while teacher is not because an individual can become and stop being a teacher through time. A concept classified as a *Kind* is a concept which supplies the identity criterion to its instances, while a *Sub-kind* is a concept which inherits its identity criterion from the kind concept. A *Quantity* is a concept in which the set of its individuals refers to portions of some substance like water, for example. *Quality dimension* is a structure used to represent the set of values (*Qualia*) associated to a rigid sortal. Each of the values from a quality dimension is a *Quale*.

The conceptual modeling process starts by selecting objects in reality and evaluating their adequacy with representing primitives. Selecting the best primitive that fits for representing an object is a key point in conceptual modeling. Thus, meta-properties and meta-concepts analysis have turned to be a powerful tool on the evaluation and selection of primitives during the ontology construction process.

3. Visual Representations

A representation is characterized by an entity assuming the role of representing another one. Furthermore, a representation is characterized by the set of relationships established between the representing entity and the represented entity [Gurr 1999]. The interpretation of a piece of knowledge represented using propositional languages is achieved by concatenating the symbols. However it is possible to explore the visual-language system intrinsic properties in order to explore the direct correspondence among the concept properties and the visual representation properties [Gurr 1999].

Atsushi Shimojima defines as inferential "free-rides" the possibility to capture semantic information through the direct correspondence among the representations' and the concepts' visual properties [Shimojima 1996]. Figure 2 depicts the use of free-rides representing two equivalents logical syllogisms. The same conclusion (iii) is achieved in a more straightforward way on Figure 2-b while exploring the correspondence between the concept and the visual representation properties.

Based on the main characteristics, Peirce *apud* [Burks 1949] classifies representations as *symbol*, *index*, and *icon*. A symbol has no direct or indirect relationship with its meaning. The meaning of a symbol is established by convention, which must be known.

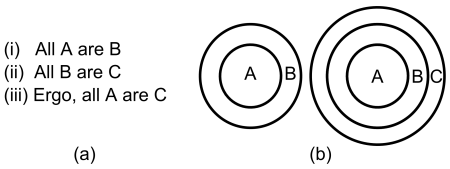


Figure 2. Equivalent logical syllogisms represented using propositional language (a) and the Euler Circles visual language (b) [Guizzardi et al. 2002].

Indexes have associative and indicative relations with their meanings. The mercury column height of a barometer is an index of the atmospheric pressure. Icons seem with what they mean. Thus, the meaning of an icon is captured by the same perception process used to recognize the originally represented object or event.

4. Meta-constructs

In order to formalize visual knowledge using domain ontologies, two hybrid meta-constructs are formally defined. They are considered hybrid because the concepts classified by them are represented by a pair consisting in one propositional and one pictorial representation. While the propositional representation formalizes the domain vocabulary and it is used for communication purposes, the pictorial representation formalizes the visual knowledge that the domain expert is not able to verbally express. The representations do not fully overlap but they complement each other. These two meta-constructs are built based on the Guizzardi's UFO and the extension of the Ullmann triangle.

PictorialConcept is the meta-construct responsible by representing visual types. Based on the meta-properties proposed by Guarino this meta-construct gives or carries an identity criterion, has ontological rigidity, and unity property. According to that, and based on the UFO, this meta-construct represents the concepts classified as rigid sortals, i.e., it represents kind, sub-kind, collective, and quantity meta-concepts. The left side of Figure 3 depicts the propositional representation of a geological sedimentary structure while the right side shows the iconic representation used to express the non-verbalizable knowledge.

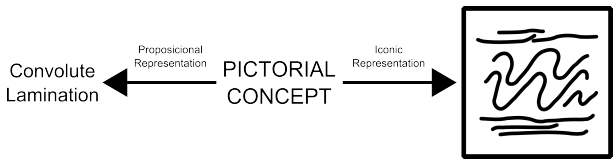


Figure 3. Application of the meta-construct Pictorial Concept.

PictorialAttribute is the meta-construct created to represent the quality dimensions' values from (visual) concept attributes. The quality dimensions classified by this meta-construct give or carry identity criterion, have ontological rigidity, but, differently from the previous meta-construct, they do not have unity criterion. Therefore, based on the UFO, this meta-construct represents concepts classified as quale, i.e., it represents the value set associated to a quality dimension. Figure 4 left side shows the propositional representation while the right side depicts the pictorial representation for the quale of the sorting attribute of sedimentary rocks.

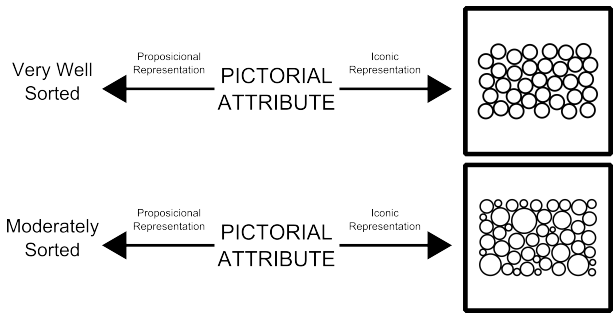


Figure 4. Application of the meta-construct Pictorial Attribute.

5. Visual Knowledge based system

We have built a stratigraphy domain ontology based on the proposed meta-constructs. Stratigraphy, an imagistic domain, is a sub-area from geology responsible to study the formation processes of sedimentary rocks [Press et al. 2004] where the geologist's analysis is based on the rock visual features. The formalized ontology represents the concepts both in Portuguese and English. Using the proposed meta-constructs the iconic vocabulary is defined for the visual types and for the visual attributes in association with the propositional vocabulary.

The visual knowledge based system, built upon the domain ontology, is used to visually describe petrological features (litology, textures, structures) in cores collected from exploration wells, that will be further used to understand the structural correlation of geological units in petroleum exploration. The interface of the system is fully based on the pictorial concepts and attributes which keeps the interaction closer to the way geologists use to describe cores. The pictorial features are, by their side, internally related to propositional concepts that give to the system the capability of extract geological correlation. Therefore, the hybrid model guarantees a unique ability to our system: depicts to the user core descriptions based on pictorial visualizations the user are used to, while capture real knowledge for further correlation and inference. The full model is persisted in a relational database as well the user data.

The knowledge base is built to be scalable in order to grow as the expert geologists use the system and find/create new (visual) concepts and (visual) attributes. Once the knowledge base is referenced by the user-data database, any change on the former is automatically reflected on the latter, ensuring the data integrity.

6. Conclusion

Domain novices take long time and lots of resources to be trained and accumulate sufficient knowledge to become a domain expert. When dealing with image-based problems, experts build their ability by accumulating internal abstract representations of the key visual features that allow solving problems in the domain. An intrinsic characteristic of the imagistic-domain experts is to use drawings to externalize and express their mental models. Imagistic-domain experts have difficulties to externalize their knowledge and make it accessible to share with others.

Integrating visual content into ontologies opens the possibility to formalize the visual knowledge from imagistic-domain experts. Our proposed meta-constructs give

the first step in that direction. One advantage of using the proposed meta-constructs to formalize visual knowledge is to make a very specialized kind of knowledge accessible to novices. Another advantage of their usage is to open the possibility of constructing visual-knowledge based systems, since the knowledge becomes machine readable. Future directions of the studies will be exploring the relationship among properties of concepts and properties of pictorial representations in order to better understand the externalization process and to create more accurate representations.

7. Acknowledges

The scholarships that supported this project were sponsored by the Government Agency of CNPq. SEBRAE and Endeeper Co. have provided financial support.

References

Abel, M., Silva, L. A. L., Campbell, J. A., and Ros, L. F. D. (2005). Knowledge acquisition and interpretation problem-solving methods for visual expertise: S study of petroleum-reservoir evaluation. *Journal of Petroleum Science and Engineering*, 47(1-2):51 – 69. Intelligent Computing in Petroleum Engineering.

Burks, A. W. (1949). Icon, index, and symbol. *Philosophy and Phenomenological Research*, 9(4):673–689.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2002). *Sweetening Ontologies with DOLCE*, chapter Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, pages 223–233. Springer Berlin / Heidelberg.

Guarino, N. and Welty, C. (2002). Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65.

Guarino, N. and Welty, C. A. (2004). An overview of ontoclean. In *Handbook on Ontologies*, pages 151–172. Springer.

Guizzardi, G. (2005). *Ontological Foundations for Structural Conceptual Models*. Enschede, The Netherlands: Universal Press, 410p. (CTIT PhD Thesis Series).

Guizzardi, G., Pires, L. F., and Sinderen, M. J. V. (2002). On the role of domain ontologies in the design of domain-specific visual modeling languages. In *Second Workshop on DomainSpecific Visual Languages, 17th Annual ACM Conference on ObjectOriented Programming, Systems, Languages, and Applications.*

Gurr, C. A. (1999). Effective diagrammatic communication: Syntactic, semantic and pragmatic issues. *Journal of Visual Languages and Computing*, 10:317–342.

Lorenzatti, A., Abel, M., Fiorini, S., Bernardes, A., and dos Santos Scherer, C. (2011). Ontological primitives for visual knowledge. In da Rocha Costa, A., Vicari, R., and Tonidandel, F., editors, *Advances in Artificial Intelligence - SBIA 2010*, volume 6404 of *Lecture Notes in Computer Science*, pages 1–10. Springer Berlin / Heidelberg.

Press, F., Siever, R., Groetzinger, J., and Jordan, T. H. (2004). *Para Entender a Terra*. Bookman, 4 edition.

Shimojima, A. (1996). Operational constraints in diagrammatic reasoning. *Logical Reasoning with Diagrams*, pages 27–48.

Ullmann, S. (1979). *An Introduction to the Science of Meaning*. Oxford.

Extração e Validação de Ontologias a partir de Recursos Digitais

Kassius Prestes[1], **Rodrigo Wilkens**[1], **Leonardo Zillio**[2], **Aline Villavicencio**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul(UFRGS)

[2]Instituto de Letras – Universidade Federal do Rio Grande do Sul(UFRGS)

Abstract. *This paper aims at presenting a methodology for semi-automatic validation of an wide-coverage ontology based on an existing electronic resource, PAPEL. From the existing relations, we choose those of synonymy and hypernymy to generate the ontology. The resulting output was converted to OWL format e manually validated by a lexicographer. As result, we have a wide-coverage ontological resource that can be used in different subareas of computer science. The resource displays concepts organized according to their hypernymy and validated synonymy relations.*

Resumo. *O objetivo desse trabalho é apresentar uma metodologia para validação semiautomática de uma ontologia de ampla cobertura, com base em um recurso eletrônico existente, o PAPEL. Das relações disponíveis, foram usadas as de sinonímia e de hiperonímia para construção da ontologia. Os resultados foram convertidos para o formato OWL e manualmente validados por um lexicógrafo. O resultado obtido foi um recurso ontológico de ampla cobertura que pode ser empregado em diversas áreas da computação. O recurso apresenta termos organizados a partir de suas relações de hiperonímia e de sinonímia, sendo estas validadas.*

1. Introdução

Sistemas computacionais como sistemas de perguntas e respostas (P&R) têm a tarefa de responder automaticamente uma questão em linguagem natural, procurando por informações em fontes de dados, tais como um banco de dados estruturado ou documentos não-estruturados em linguagem natural (p. ex., jornais). Esse tipo de sistema normalmente realiza quatro passos: análise da pergunta; identificação dos documentos candidatos; geração das respostas candidatas; e pontuação das respostas. Exemplos são [Kaisser 2005], [Lin 2005], [Zheng 2002], [Sarmento et al. 2008] e [Amaral et al. 2006].

Um exemplo de sistema de P&R para o português é o projeto Comunica [Wilkens et al. 2010], que busca responder perguntas sobre transferências constitucionais de municípios via telefone. Nele, tanto a pergunta do usuário quanto a resposta do sistema são em liguagem natural, visando a uma maior inclusão digital. O projeto se divide em quatro módulos-chave: reconhecimento de voz, processamento de texto, acesso a banco de dados e síntese de voz. O módulo de reconhecimento de voz realiza a conversão de áudio para texto. O processamento de texto tem a função de identificar dados relevantes informados pelo usuário a partir da frase transcrita (pelo módulo de reconhecimento de voz). A identificação dos conceitos é feita por meio de duas ontologias que validam as palavras da pergunta do usuário: uma de propósito geral e uma do domínio da aplicação. Os conceitos identificados são então buscados pelo módulo de acesso a banco de dados e a resposta gerada é sintetizada pelo módulo de síntese de voz. Para tal aplicação, é

necessária uma ontologia de ampla cobertura que possa auxiliar tanto o módulo de reconhecimento de voz a validar as palavras reconhecidas quanto o módulo de processamento de texto na identificação de conceitos.

O objetivo deste trabalho é apresentar a metodologia de extração e validação da ontologia de propósito geral usada no âmbito do projeto Comunica. A ontologia foi extraída de modo semiautomático a partir do PAPEL (Palavras Associadas Porto Editora Linguateca) [Oliveira et al. 2008] e convertida para o formato OWL. Os resultados da conversão foram validados por um lexicógrafo. O PAPEL é um conjunto de relações entre palavras extraídas automaticamente das definições em um tesauro eletrônico. Ele contém 199.672 entradas, distribuídas em 8 tipos de relações. Dentre essas, foram selecionadas as relações de hiperonímia [1] e de sinonímia [2] como base para a ontologia. Neste trabalho, são descritos os processos de conversão dos dois tipos de relações, com uma discussão da validação das relações de sinonímia.

Este artigo é estruturado da seguinte maneira: na Seção 2, são discutidos alguns trabalhos relacionados; a metodologia para extração e validação é descrita nas Seções 3 e 4; por fim, a Seção 5 apresenta conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Buitelaar [Buitelaar et al. 2005] define o aprendizado de ontologias como a aquisição de conhecimento a partir de textos. Grande parte desse trabalho utiliza como base áreas da computação como processamento de linguagem natural, inteligência artificial e aprendizado de máquina. Existem diversas abordagens para a extração automática de ontologias. Algumas abordagens são probabilísticas, como em [Venant 2008] [Chotimongkol and Rudnicky 2002] [Salton and Buckley 1988]. Contudo, uma das abordagens mais exploradas é a identificação de relações a partir de expressões textuais, como mostrado em [Hearst 1992]. A identificação baseada em expressões apresenta um melhor resultado na extração de documentos que possuem uma estrutura mais ou menos fixa, como dicionários, por isso, essa abordagem foi utilizada para a criação do PAPEL.

Ontologias, especialmente as de ampla cobertura, são recursos de grande valor para sistemas que empregam tecnologias de linguagem. Por exemplo, a WordNet [Miller 1995] é um recurso eletrônico que contém relações semântico-conceptuais e lexicais entre as palavras. Ela foi originalmente desenvolvida pela Universidade de Princeton para o inglês, e posteriormente estendida para outras línguas, inclusive para o português [Marrafa et al. 2005]. Nela, os termos são agrupados em synsets, onde todos os sinônimos de um termo estão no mesmo grupo que ele, contendo uma definição e um conjunto de relações linguísticas. A Wordnet é utilizada em aplicações como tradução automática, sistemas de busca e extração de informação, entre outros. A WordNet do português (WordNet.PT [3]) contém cerca de 19.000 termos, distribuídos em vários campos semânticos. O fragmento disponível é composto por termos de diversos domínios, como arte, saúde, transportes e vestuário.

[1] hiperonímia é uma relação entre palavras que dá idéia de um todo, da qual se originam diversas ramificações, por exemplo, veículo é hiperônimo de carro, barco e avião

[2] sinonímia é uma relação entre palavras da mesma categoria gramatical, com sentido parecido e com forma diferente, como por exemplo carro e automóvel

[3] Desenvolvida pelo Centro de Linguística da Universidade de Lisboa pelo CLG - Grupo de Computação do Conhecimento Léxico-Gramatical.

Outro importante recurso léxico para o português é o PAPEL, desenvolvido com o objetivo de prover uma ontologia geral da linguagem, de grande abrangência [Oliveira et al. 2008]. Esse recurso foi construido através de extração semiautomática baseada em padrões de expressões que ocorrem nas definições do Dicionário da Língua Portuguesa [Editora 2005]. Dessa forma, foram identificadas relações composicionais, hierárquicas e de sinonímia. Exemplos dessas relações seriam:

repartir SINONIMO_DE partilhar
vasqueiro PROPRIEDADE_DE_ALGO_QUE_CAUSA vasca
vazar ACCAO_QUE_CAUSA vazão
cabo PARTE_DE vassoura
navio HIPERONIMO_DE veleiro

Devido à sua abrangência (com 199.672 entradas), foi realizada uma avaliação por amostragem dos resultados da extração semiautomatica [Oliveira et al. 2009], sendo que 50% das relações de sinonímia apresentam erros em potencial. Por exemplo: deliberadamente SINONIMO_DE peito. Para contornar estes problemas apresentamos uma metodologia de extração e validação mais confiáveis (pela validação manual) inseridas no projeto Comunica[Wilkens et al. 2010].

3. Metodologia

Para a construção de uma ontologia de alta cobertura e precisão a partir dos dados disponibilizados no PAPEL, foi necessária a definição de uma metodologia para conversão e validação sistemática do recurso com a construção de um sistema para identificação automática de conflitos nas entradas definidas. Neste trabalho, é abordada a conversão de duas relações para o format OWL: sinônimos e hiperônimos.

3.1. Relações de Hiperonímia

O PAPEL contém 61.263 entradas com relações de hiperonímia expressas no seguinte formato: palavra_1 HIPERONIMO_DE palavra_2. Para a conversão, foi utilizada a linguagem de programação Java e o framework Jena. A metodologia de conversão prevê duas etapas: conversão passiva e conversão ativa.

3.1.1. Conversão Passiva

A abordagem passiva consiste em armazenar as classes ontológicas apenas quando estas são apresentadas pelo PAPEL (as classes vão sendo criadas enquanto o arquivo de relações é lido). Para cada entrada do PAPEL, definida em termos de duas palavras (palavra_1 e palavra_2):

1. Para cada palavra:
   (a) Verificar se a palavra já existe como classe (dadas as várias palavras repetidas no PAPEL),
   (b) Se não existe, criar a classe.
2. Adicionar à ontologia a relação onde a classe palavra_1 é superclasse da classe palavra_2.

Esse processo é repetido para cada uma das relações de hiperonímia existentes. Devido ao número de entradas do PAPEL e às várias comparações necessárias, esta abordagem se mostrou muito custosa em termos de processamento e alto consumo de memória [4].

[4] Apenas cerca de 3% das relações em um período de aproximadamente 24 horas.

3.1.2. Conversão Ativa

Para tornar o processo mais eficiente em ontologias de alta cobertura, a segunda abordagem proposta cria inicialmente uma nova ontologia esquemática, e as informações já validadas somente são inseridas na nova ontologia ao final de cada passo de validação. Esse processo é composto por duas etapas. Primeiro, define-se uma ontologia básica com todas as classes necessárias, mas sem as relações entre elas. Para tanto, extraem-se do PAPEL todas as palavras sem repetição e cria-se uma classe para cada uma delas. A partir dessa ontologia básica com as classes necessárias, adicionam-se as relações entre as classes. Dada uma definição no PAPEL no formato *palavra_1 HIPERONIMO_DE palavra_2*, procura-se na ontologia básica as classes relativas a palavra_1 e palavra_2 e adiciona-se a relação à ontologia. Como resultado, obtem-se a conversão da estrutura de hiperonimos do PAPEL para o formato OWL.

3.2. Relações de Sinonímia

A base de sinônimos do PAPEL possui relações expressas da seguinte maneira: palavra_1 SINONIMO_<classe>_DE palavra_2, onde em <classe> ocorrem as seguintes tags, que indicam classes gramaticais: N substantivo, V verbo, ADJ adjetivo e ADV advérbio. Dado o contexto deste trabalho, foram extraídas as relações de sinonímia entre substantivos [5]. A extração de sinônimos foi realizada através das seguintes etapas:

1. Criar uma lista (inicialmente vazia) de conjuntos (inicialmente vazios). Cada conjunto armazenará palavras que são sinônimas entre si.
2. Para cada uma das entradas, identificar as duas palavras sinônimas presentes.
3. Verificar se uma dessas palavras já se encontra em algum conjunto de sinônimos existente.
    (a) Se sim, insere-se a outra palavra nesse conjunto.
    (b) Se as palavras não estavam em nenhum conjunto existente, cria-se um novo conjunto contendo ambas.

Dada a ampla abrangência do PAPEL, a polissemia das palavras e a transitividade da relação de sinonímia (se A é sinônimo de B e B é sinônimo de C, então A é sinônimo de C), ao final do processo, quase todas as palavras foram reconhecidas como sinônimas entre si. Para criar uma ontologia de alta precisão e cobertura, a aplicação da metodologia foi semiautomática, e esses casos foram supervisionados por um lexicógrafo, com a seguinte modificação no passo 3:

3. Verificar se uma dessas palavras já se encontra em algum conjunto de sinônimos existente.
    (a) Se sim, realizar a verificação manual da ambiguidade, analisando os dois conjuntos em que as palavras seriam inseridas. Caso necessário, editar os conjuntos manualmente, separando as palavras de modo adequado.
    (b) Se nenhuma estava em um conjunto existente, criar um novo conjunto com ambas.

Ao final do processo obtem-se um amplo conjunto de grupos de sinônimos, que pode ser integrado à estrutura da ontologia gerada no passo anterior.

[5] Porém, essa abordagem pode ser, em princípio, aplicada diretamente aos outros tipos de relação.

4. Validação das Relações de Sinonímia

As relações de sinonímia da ontologia resultante foram manualmente validadas. Essa validação ocorreu com consulta a dicionários de língua portuguesa e a contextos reais de ocorrência dos pares de sinônimos analisados [6]. Caso as definições dos dicionários não aclarassem o problema, utilizou-se o buscador on-line do Yahoo! para se observarem também os contextos de ocorrência.

Dada a magnitude do recurso gerado, para palavras polissêmicas, a validação foi realizada com base no significado mais frequente apropriado para o grupo de sinônimos, sendo que cada substantivo poderia estar presente em apenas um grupo de sinônimos. Por exemplo: dada a avaliação da relação de sinonimia proposta entre abatimento, diminuição e desânimo, apesar de desânimo e diminuição não parecerem sinônimas sem um contexto muito específico, a palavra abatimento pode ser considerada sinônima de ambas, entre outras. A consulta a dicionários retornou informações sobre abatimento de animais, abatimento de preços (diminuição) e abatimento emocional (desânimo). No buscador do Yahoo!, entre as primeiras 20 ocorrências de "abatimento", havia 12 ocorrências "abatimento de preços", 4 de "abatimento emocional" e 1 de "abatimento de animais"; as outras eram irrelevantes (definições de dicionários on-line etc.). Como resultado, abatimento foi incluída com diminuição e excluída do conjunto de desânimo.

A adoção da metodologia proposta permite reduzir a subjetividade envolvida em todo o processo de decidir que palavras são sinônimas entre si e quais não devem ser [7]. Isso se torna importante para a replicabilidade do processo de validação, tendo em vista a natureza inerentemente subjetiva e depedente do vocabulário do avaliador da decisão de quais palavras possuem uma semelhança de significado.

Ao final desse processo, a ontologia resultante contém 46.904 entradas, com 40.614 relações de hiperonímia e 20.096 de sinonímia. O recurso apresenta alta abrangência e pode ser utilizado em uma grande variedade de sistemas de tecnologia de linguagem.

5. Conclusões e Trabalhos Futuros

Este artigo propôs uma metodologia para validação de uma ontologia com relações de hiperonímia e sinonímia a partir de um recurso lexical eletrônico. Foi apresentada em detalhes a avaliação de sinonímia realizada manualmente e com decisões auxiliadas por outros recursos. Apesar da magnitude do recurso original e das relações a serem adicionadas, essa metodologia possibilitou uma validação mais ampla do recurso, em vez da avaliação por amostragem proposta em [Oliveira et al. 2009]. Após a etapa apresentada neste trabalho, resta ainda a validação da estrutura de hiperônimos. Contudo, a parte já validada do recurso permite a sua utilização em diversas áreas da computação, tais como tradução automática, sistemas de busca e extração de informação, sistemas conversacionais, sistemas de inferência e extração automática de ontologias.

Agradecimentos

Esta pesquisa tem apoio dos projetos COMUNICA (FINEP/SEBRAE 1194/07), CAPES-COFECUB (707/11) e CNPq (479824/2009-6 e 309569/2009-5).

[6] Os dicionários utilizados foram o Houaiss Eletrônico (disponível em CD), o Michaelis On-line e o Dicionário da Língua Portuguesa da Porto Editora.

[7] Tal procedimento pode ser suficiente para os casos que envolvem mais de um significado frequente, mas, dada a abrangência do recurso, tal avaliação manual se torna impraticável.

References

Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., and Pinto, C. (2006). Priberams question answering system for portuguese. *CLEF*.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12.

Chotimongkol, A. and Rudnicky, A. (2002). Automatic concept identification in goal-oriented conversations. In *Seventh International Conference on Spoken Language Processing*.

Editora, P. (2005). *Dicionário PRO da Língua Portuguesa*. Porto.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Kaisser, M. (2005). Qualim at trec 2005: Web-question answering with framenet. *TREC*.

Lin, J. (2005). Evaluation of resources for question answering evaluation. Technical report, University of Maryland, College Park.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2005). Wordnet.pt uma rede léxico-conceptual do português on-line. *XXI Encontro da Associação Portuguesa de Linguística*, pages 28–30.

Miller, G. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Oliveira, H. G., Santos, D., and Gomes, P. (2009). Extracção de relações semânticas entre palavras a partir de um dicionário: o papel e sua avaliação. *STIL 2009, Linguamática*, pages 77–93.

Oliveira, H. G., Santos, D., Gomes, P., and Seco, N. (2008). Papel: A dictionary-based lexical ontology for portuguese. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Proceedings of Computational Processing of the Portuguese Language (PROPOR)*, volume 5190 of *LNAI*, pages 31–40. Springer.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval* 1. *Information processing & management*, 24(5):513–523.

Sarmento, L., Teixeira, J. F., and Oliveira, E. (2008). Experiments with query expansion in the raposa (fox) question answering system. *The Cross-Language Evaluation Forum (CLEF)*.

Venant, F. (2008). Semantic visualization and meaning computation. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 185–188. Association for Computational Linguistics.

Wilkens, R., Villavicencio, A., Muller, D., Wives, L., da Silva, F., and Loh, S. (2010). Comunica - a question answering system for brazilian portuguese. *Coling 2010*.

Zheng, Z. (2002). Answerbus question answering system. *Proceeding of HLT Human Language Technology Conference (HLT 2002)*.

Sistema de Aquisição semi-automática de Ontologias

Gabriel Gonçalves[1], **Rodrigo Wilkens**[1], **Aline Villavicencio**[1,2]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (Brasil)

[2]CSAIL, MIT (EUA)

gabrielgonc@gmail.com, {rwilkens,avillavicencio}@inf.ufrgs.br

Abstract. *This paper presents an ongoing work on ontology learning from text, focusing on the acquisition of concepts and relations. In order to do that, this work investigates approaches for ontology learning, and presents a proposal based on graphs metrics to identify concepts, and text analysis to find relations between the concepts.*

Resumo. *Este artigo apresenta um trabalho em andamento na área de aprendizado de ontologias a partir de texto, focando na identificação de conceitos e relações. Para isto, este trabalho investiga abordagems para o aprendizado de ontologias e apresenta uma proposta baseada métricas de grafos para identificar conceitos, e análise do texto com os conceitos encontrados para obter relações.*

1. Introdução

Em alguns sistemas computacionais como sistemas de perguntas e repostas e agentes conversacionais, para suprir as necessidades de informações de usuários, pode ser necessário utilizar informações não-estruturadas, como as disponíveis na web, e realizar um processamento dessas informações. Para tanto, diversas linguagens e padrões vem sendo desenvolvidos, tais como *Resource Description Framework* [3] e *Web Ontology Language* [1], que permitem a definição de conceitos e a descrição de suas relações e propriedades. Segundo o W3C (World Wide Web Consortium) [13], para sistemas que precisam compartilhar conhecimentos do mesmo domínio (por exemplo, medicina, mercado imobiliário e petróleo) é necessário o uso de ontologias para unificar este conhecimento. Contudo, o processo de criação de ontologias de forma manual é custoso em termos de tempo e recursos e exige um especialista do domínio. Desta forma, algumas tarefas desse processo tem sido automatizadas em sistemas computacionais, como mostrado em [16], [18], [11] e [6].

Em geral o aprendizado automático de ontologias é visto como a aquisição de conhecimento a partir de textos, onde grande parte do trabalho utiliza como base áreas da computação como processamento de linguagem natural, inteligência artificial e aprendizado de máquina [2]. Para Yang e Jamie [18] o processo de construção de ontologias ocorre em quatro passos: (1) detectar candidatos a conceitos; (2) agrupar conceitos similares; (3) encontrar um nome para cada grupo; (4) formar uma árvore para representar a ontologia.

Para muitas línguas e domínios o aprendizado de ontologias tem que ser realizado a partir de poucos recursos linguísticos disponíveis. Nesse contexto, este trabalho objetiva

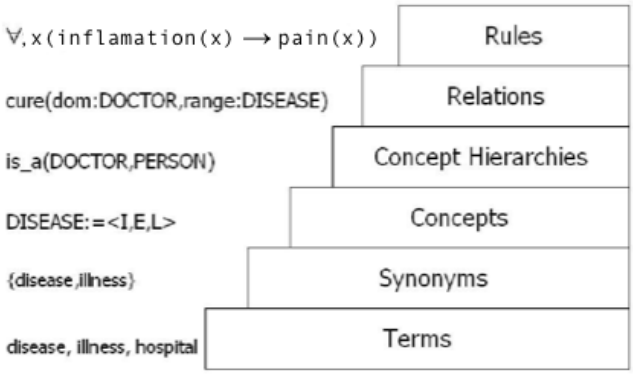


Figura 1. Hierarquia dos processos de aprendizado de ontologia [2]

investigar dois aspectos do aprendizado de ontologias, a identificação de conceitos e de relações entre conceitos, focando na identificação de conceitos simples e na identificação de elementos que indicam relações entre termos. Para tanto esse trabalho inicia com uma revisão do estado da arte, na seção 2. A seguir, na seção 3 são apresentadas as técnicas utilizados na abordagem proposta. Na seção 4 são discutidas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Gruber [10] define uma ontologia como uma especificação formal e explícita de uma conceitualização compartilhada por um domínio de interesse, onde formal significa que a ontologia deve ser interpretável por computador e aceita por um grupo ou comunidade da área que a ontologia modela. Além disso, deve ser restrita a um dado domínio de interesse e, portanto, modelar conceitos e relações relevantes a uma tarefa ou aplicação particular do domínio [2]. Atualmente não há um consenso sobre os métodos para o aprendizado automático de ontologias, que segundo [2], podem ser divididos em seis níveis: termos, sinônimos, conceitos, hierarquias de conceitos, relações e regras. A hierarquia dessas tarefas no processo de aprendizado de ontologias é mostrada na Figura 1.

A aquisição de termos consiste em encontrar automaticamente palavras que representem conceitos de um domínio. Este é o passo inicial do aprendizado de ontologias, sendo seus resultados usados em todas as etapas posteriores. As técnicas mais utilizadas para tanto são a indexação de termos, análise de frequência, coocorrência e uma combinação dos dois métodos anteriores [14]. Segundo Buitelaar [2], a extração de conceitos é uma etapa controversa, por não estar claro o que exatamente é um conceito. Nesta etapa podem ser considerados como conceitos uma definição, instancias de um conceito ou um conjunto multilíngue de termos, dependendo do uso que o pesquisador da ontologia gerar.

A identificação de sinônimos visa a aquisição semântica de variantes de termos, ou seja, encontrar entre os termos de um texto aqueles que compartilham funções semânticas. Para tanto, o estado da arte mapeia a semântica de cada palavra e identifica as palavras que possuem intersecção, sendo este mapeamento comumente realizado pelo contexto dos termos [3] ou diretamente pala semântica dos termos [17].

A extração de taxonomias busca identificar uma organização hierárquica entre

os conceitos, sendo comum o uso de listas de termos que indicam tais relações, o que gera uma boa precisão na identificação, mas devido ao fato destes padrões serem muito específicos esta abordagem apresenta uma baixa cobertura das relações existentes [11]. Outra abordagem é a hipótese de distribuição, onde são derivadas automaticamente as hierarquias de termos a partir do texto usando análise de conceitos formais [8] (ex. [4], [7], [9]). A comunidade de recuperação de informação trata esta tarefa a partir da avaliação da distribuição e relevância dos termos nos documentos, como mostrado por Sanderson e Croft em [15].

A extração de outras relações não hierárquicas entre conceitos (por exemplo, relações entre sintomas, doenças e drogas) tem sido feita a partir de textos, em geral procurando por relações entre pares de conceitos com mesma classe gramatical.

Por fim, a extração de regras, discutida em [12] e [5], é a área pesquisada menos abordada em aprendizado de ontologias [2]. O objetivo deste passo é encontrar regras gramaticais que rejam as relações das ontologias.

Dentro desse contexto, esse trabalho é similar ao de [3] no uso de mutual information para a extração de sinonimia, com a difierença de que utilizamos esta métrica sobre um grafo do texto, e não diretamente sobre ele, e a [16] que verificam relações, diferindo por generalizarmos os padrões encontrados.

3. Metodologia

O objetivo deste trabalho é gerar automaticamente ontologias a partir de um corpus do domínio, com foco na identificação de conceitos e relações do domínio, discutidos respectivamente nas seções 3.1 e 3.2.

3.1. Aquisição de Termos e Conceitos

Neste trabalho não diferenciamos termos e conceitos no processo de aquisição devido à natureza próxima destes, assim tornando o resultante do sistema mais próximo de uma ontologia linguística de domínio. O processo inicia com a geração de um grafo a partir do corpus, onde as palavras são os nós, que são ligados uns aos outros quando as palavras que formam os nós encontram-se na mesma sentença, como ilustrado na Figura 2. Nas Figuras 2.i e 2.ii, as frases "João e Maria foram ao parque domingo"e "Domingo o parque estava lotado", respectivamente, são transformadas em grafos. As duas frases unidas geram um grafo, cujas arestas são pesadas de acordo com o número de vezes que cada par de nós coocorre no texto. (Figura 2.iii). Sobre este grafo utilizamos as seguintes métricas de grafos para gerar candidatos a conceitos:

- **centralidade** para verificar a importância do nó no grafo,
- **grau**, que representa o numero de ligações de um nó e
- ***closeness***, que verifica a média dos caminhos mínimos para se chegar ao nó.

3.2. Aquisição de Relações

Para a obtenção das relações não hierárquicas realizamos uma análise do corpus para identificar possíveis expressões que indiquem alguma relação entre os termos. Este processo foi dividido em três etapas sequenciais: extração de relações; generalização das relações para obter padrões; e re-extração das relações utilizando os padrões encontrados.

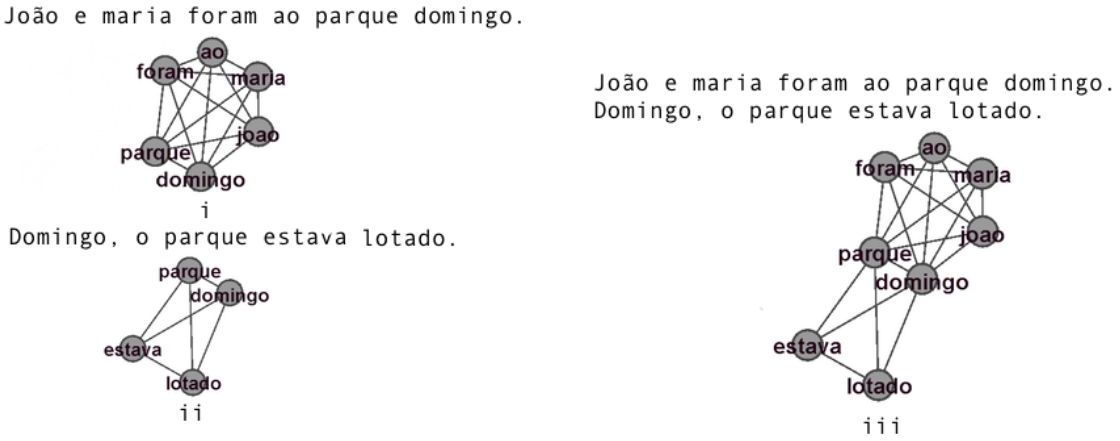


Figura 2. Exemplo de texto transformado em grafo.

Para a extração de relações o sistema identifica no corpus todos os conceitos e segmenta as palavras que ocorrem entre eles.[1] Todas as palavras que se encontram entre um par de conceitos são consideradas candidatas a relação. Estas relações candidatas são filtradas, permanecendo apenas palavras cujas classes gramaticais são permitidas (neste ponto utilizamos filtros que combinam informações lexicais e morfosintáticas para uma extração mais direcionada). Desta forma é obtida a primeira lista de relações entre conceitos (este processo é exemplificado na Figura 3, onde duas relações distintas são encontradas para a frase[2] entre os conceitos *obras* e *licensas*, e *distribuição* e *trabalhos*).

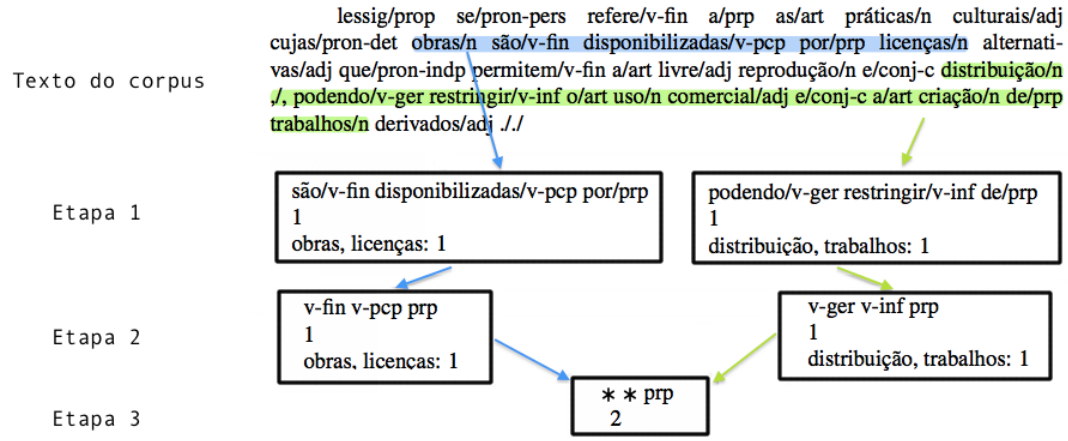


Figura 3. Extração de relações entre conceitos.

Na segunda etapa, generalização das relações, consideramos as relações apenas como uma sequência de classes gramaticais (no exemplo da Figura 3, etapa 2, as palavras são substituídas por suas classes gramaticais). Estas relações formam uma segunda lista, onde estão as relações compostas de classes gramaticais e suas respectivas frequências.

[1] Assume-se que não pode haver um conceito entre um par de conceitos.

[2] A frase está anotada com suas classes gramaticais (prop: nome próprio, pron-pers: pronome pessoal, v-fin: verbo finito, prp: preposição, art: artigo, adj: adjetivo, pron-det: pronome determinado, n: substantivo, v-pcp: verbo no particípio, pron-ind: pronome indeterminado, conj-c: conjunção coordenada, v-ger: verbo no gerúndio, v-inf: verbo no infinitivo).

Neste ponto, as relações são generalizados de acordo com seu número de palavras e de classes gramaticais que compartilham a mesma posição. Na Figura 3, etapa 3, as duas relações têm o mesmo tamanho e compartilham o mesmo elemento na posição três, gerando uma nova relação genérica contendo três elementos, restringindo apenas o terceiro.

O objetivo da primeira etapa é mostrar as relações que ocorrem diretamente no corpus, enquanto a segunda etapa objetiva criar padrões genéricos de identificação. Com estas informações, a terceira etapa, re-extração das relações, utiliza a lista gerada pela etapa 2 como modelo para identificar novas relações no corpus, ou seja, relações que não foram identificadas na primeira etapa.

4. Conclusões e Trabalhos Futuros

O aprendizado de ontologias é um campo interdisciplinar, que abrange diversas áreas da computação, como processamento de linguagem natural. As propostas para aprendizado semi-automático de ontologias permitem diminuir consideravelmente o custo e esforço envolvidos na construção de ontologias.

Dentro desse contexto, esse trabalho apresentou uma abordagem baseada em grafos para a identificação de termos e relações a partir de corpora. Essa abordagem permite extrair de forma recursiva novas expressões que PODEM indicar relações entre termos.

Como trabalhos futuros se prevê uma avaliação sistemática dos resultados obtidos, por cada etapa do processo, por um especialista do domínio. Os trabalhos futuros envolvem ainda a aquisição de sinônimos e aquisição de relações hierárquicas, assim permitindo alem da identificação das relações gerais, aquelas relações mais específicas (por exemplo, “tipo de”, “é um”). Pretendemos também validar os resultados obtidos com o sistema utilizando corpus de diferentes domínios, como o corpus GENIA [3] do domínio de biologia.

Agradecimentos

Esta pesquisa tem apoio dos projetos COMUNICA (FINEP/SEBRAE 1194/07), CAPES-COFECUB (707/11) e CNPq (479824/2009-6, 202007/2010-3 e 309569/2009-5).

Referências

[1] Resource description framework (rdf) model and syntax, 2011.

[2] P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12, 2005.

[3] A. Chotimongkol and A.I. Rudnicky. Automatic concept identification in goal-oriented conversations. In *Seventh International Conference on Spoken Language Processing*, 2002.

[4] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305–339, 2005.

[5] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. *Machine Learning Challenges*, pages 177–190, 2006.

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus/

[6] E. Drymonas. Ontology learning from text based on multi-word term concepts: The ontogain method. Master's thesis, Department of Electronic and Computer Engineering, Technical University of Crete, Greece, 2009.

[7] D. Faure and C. Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, pages 707–728. Citeseer, 1998.

[8] B. Ganter and R. Wille. Formal concept analysis. *WISSENSCHAFTLICHE ZEITSCHRIFT-TECHNISCHEN UNIVERSITAT DRESDEN*, 45:8–13, 1996.

[9] G. Grefenstette. *Explorations in automatic thesaurus discovery*. Springer, 1994.

[10] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43(5):907–928, 1995.

[11] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[12] D. Lin and P. Pantel. Discovery of inference rules from text, April 5 2001. US Patent App. 09/826,355.

[13] D.L. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10:2004–03, 2004.

[14] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval 1. *Information processing & management*, 24(5):513–523, 1988.

[15] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213. ACM, 1999.

[16] F.M. Suchanek, G. Ifrim, and G. Weikum. Leila: Learning to extract information by linguistic analysis. In *Proceedings of the ACL-06 Workshop on Ontology Learning and Population*, pages 18–25, 2006.

[17] F. Venant. Semantic visualization and meaning computation. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 185–188. Association for Computational Linguistics, 2008.

[18] H. Yang and J. Callan. Metric-based ontology learning. In *Proceeding of the 2nd international workshop on Ontologies and information systems for the semantic web*, pages 1–8. ACM, 2008.

An $\mathcal{ALC}$ Description Logic Connection Method

Fred Freitas

Informatics Center - Federal Universidade of Pernambuco (CIn - UFPE)
Av. Prof. Luis Freire, s/n, Cidade Universitária, 50740-540, Recife – PE, Brazil

fred@cin.ufpe.br

Abstract. *The connection method earned good reputation in the field of automated theorem proving for around three decades, due to its simplicity, clarity, efficiency and parsimonious use of memory. This seems to be a very appealing feature, in particular in the context of Semantic Web, where it is assumed that the knowledge bases might be of arbitrary size. In this paper, I present a connection method especially tailored to infer over the description logic (DL) $\mathcal{ALC}$. Our $\mathcal{ALC}$ connection method is formalized in sequent style, although matrices should be employed for practical reasons.*

1. Introduction

The problem of reasoning over ontologies written in Description Logic (DL) [Baader et al 2003] has been receiving strong interest from researchers, particularly since the Semantic Web inception. Regarding this issue, the use of memory is certainly one important asset for a good reasoning performance. I am proposing a formalized inference system which seems adequate to address understandability and the use of a small amount of memory. Our inference system is based on the connection method (CM) [Bibel 1987], which is a simple, clear and effective inference method that has been used successfully over first order logic (FOL). Its main features clearly meet with the demands: (i) it keeps only one copy of each logical sentence in memory; and (iii) it does not derive new sentences from the stored ones.

Definition 1 (Disjunctive normal form (DNF), clause, positive matricial form). A *formula in DNF* is a disjunction of conjunctions, being in the form $C_1 \vee \ldots \vee C_n$, where each C_i is a *clause* (or *dual clause).* Clauses are conjunctions of literals like $L_1 \wedge \ldots \wedge L_m$, also denoted as $\{L_1, \ldots, L_m\}$. Formulae can be also expressed in *disjunctive clausal form* as $\{C_1, \ldots, C_n\}$. Formulae stated this way are also in *positive matricial form*, since they can be represented as a matrix. In the matrix, each clause occupies a column.

Definition 2 (Skolemization). Instead of existential quantifiers, universal quantifiers ($\forall$) are replaced by constants or Skolem functions, since I will work with the whole knowledge based negated (see next section). Variables in the resulting DNF are then (implicitly) existentially quantified.

I started with the Description Logic $\mathcal{ALC}$ (Attributive Concept Language with Complements) [Baader et al 2003], since it constitutes the foundations of many other DLs. I now present an $\mathcal{ALC}$ normal form and the $\mathcal{ALC}$ CM calculus.

2. An $\mathcal{ALC}$ Positive Matricial Normal Form

To reach this normal form, the first two actions to be made over the axioms are: (i) splitting equivalence axioms of the form $C \equiv D$ into two axioms $C \sqsubseteq D$ and $D \sqsubseteq C$, and

(ii) converting all the axioms into a *Negated Normal Form* (NNF), in which negations occurs only on literals [Baader et al 2003]. Next, I define the normal form and impurities with regard to it.

Definition 3 ($\mathcal{ALC}$ disjunction, $\mathcal{ALC}$ conjunction). An $\mathcal{ALC}$ disjunction is either a literal, a disjunction $E_0 \sqcup E_1$ or an universal restriction $\forall r.E_0$. An $\mathcal{ALC}$ conjunction is either a literal, a conjunction $E_0 \sqcap E_1$ or an existential restriction $\exists r.E_0$. E_0 and E_1 are arbitrary concept expressions.

Definition 4 ($\mathcal{ALC}$ pure disjunction). The set S_D of $\mathcal{ALC}$ pure disjunctions is the smallest set where: (i) $D_0 \in S_D$ for every literal D_0; (ii) If $D_0, D_1 \in S_D$, then $D_0 \sqcup D_1 \in S_D$; and (iii) if $D_0 \in S_D$ then $\forall r.D_0 \in S_D$. An element $\breve{D} \in S_D$ is an $\mathcal{ALC}$ pure disjunction. An $\mathcal{ALC}$ *non-pure disjunction* is an $\mathcal{ALC}$ disjunction that is not pure.

Definition 5 ($\mathcal{ALC}$ pure conjunction). The set S_C of $\mathcal{ALC}$ pure conjunctions is the smallest set where: (i) $C_0 \in S_C$ for every literal C_0; (ii) if $C_0, C_1 \in S_C$ then $C_0 \sqcap C_1 \in S_C$; and (iii) if $C_0 \in S_C$ then $\exists r.C_0 \in S_C$. An element $\hat{C} \in S_C$ is an $\mathcal{ALC}$ pure conjunction. An $\mathcal{ALC}$ non-pure conjunction is an $\mathcal{ALC}$ conjunction that is not pure.

Definition 6 (Impurity of a non-pure expression). Impurities of non-pure $\mathcal{ALC}$ DL expressions are either conjunctive expressions in a non-pure disjunction or disjunctive expressions in a non-pure conjunction. The set of impurities is called $\mathcal{ALC}$ *impurity set*, and is denoted by S_I.

Example 1 (Impurities on non-pure expressions).

The expression $(\forall r.(D_0 \sqcup ... \sqcup D_n \sqcup (C_0 \sqcap ... \sqcap C_m) \sqcup (A_0 \sqcap ... \sqcap A_p))$, a non-pure disjunction, contains two impurities: $(C_0 \sqcap ... \sqcap C_m)$ and $(A_0 \sqcap ... \sqcap A_p)$.

Definition 7 (Positive normal form). An $\mathcal{ALC}$ axiom is in positive normal form iff it is in one of the following forms: $(i)\hat{C} \sqsubseteq \breve{D}$; $(ii)C \sqsubseteq \exists r.\hat{C}$; and $(iii)\ \forall r.\breve{D} \sqsubseteq C$; where C is a concept name, $\hat{C}$ a pure conjunction and $\breve{D}$ a pure disjunction.

[Freitas et al 2011] contains $\mathcal{ALC}$ transformation algorithms to this normal form.

2.1. Translation Rules for the normalization

With all axioms in normal form, it is easy to map them both to FOL and to the matricial form, by applying the rules given in Table 1. Table 2 brings the mapping treatment of recursive sub-cases of existential and universal restrictions, when they occur inside any of the three normal forms. An improvement of the approach is, as the usual DL notation, we do not need variables, since all relations are binary.

In order to prove $KB \models \alpha$, the whole knowledge base *KB* is negated during this transformation, once we wish to prove $\neg KB \vee \alpha$ valid. Because of that, subsumption axioms of the form $C \sqsubseteq D$, which are logically translated as $C \rightarrow D$, because negated ($\neg(C \rightarrow D)$, indeed), are now translated to $C \wedge \neg D$, instead of $\neg C \vee D$. Moreover, to establish a uniform set of rules to apply over formulae, we deal with $\neg\alpha$ instead of α, so we consider formulae as $\neg A_1 \vee \ldots \vee \neg A_n \vee \neg\neg\alpha$ where $A_i \in \mathcal{T}$ (axioms in the TBox). The translation rules can then be applied over $\neg\alpha$ and all A_i.

Regarding skolemization, one representational advantage of the approach resides in the clearer matrix representation of universally quantified roles' ($\forall r.C$ or in the matrices, the negated $\exists r.C$). This construct, by definition, has the interpretation $(\forall r.C)^I = \{\forall b, (a,b) \in R^I \rightarrow b \in C^I\}$. Hence, for an axiom of the form $A \equiv \forall r.C$, the definition does not oblige concept A to dispose of instances – this is indeed a very

common error from DL users. But maybe it is not their fault: for instance, tableaux proofs over such axioms don't stress this semantics, in the sense that it allows instances of A without any role instances from r associated to it. In the $\mathcal{ALC}$ CM, the matricial representation explicits this situation: either there are no role instances ($\neg r$) or when it has a role instance (a,b), b has to be an instance of concept C.

Table 1. Translation rules to map $\mathcal{ALC}$ into FOL positive NNF and matrices.

Axiom type	FOL Positive NNF mapping	Matrix
$C \sqsubseteq \exists r.\hat{C}$, where $\hat{C} = \bigcap_{i=1}^{n} A_i$, with $A_i \in S_C$ (pure conjunction)	$(C(x) \wedge \neg r(x,f(x))) \vee (C(x) \wedge \neg A_1(f(x))) \vee \ldots \vee (C(x) \wedge \neg A_n(f(x)))$	$\begin{bmatrix} C & C & \cdots & C \\ \neg r & \neg A_1 & \cdots & \neg A_n \end{bmatrix}$
$\forall r.\breve{D} \sqsubseteq C$, where $\breve{D} = \bigcup_{j=1}^{m} A'_j$, with $A'_j \in S_D$ (pure disjunction)	$(\neg r(x,f(x)) \wedge \neg C(x)) \vee (\neg A'_1(f(x)) \wedge \neg C(x)) \vee \ldots \vee (\neg A'_m(f(x)) \wedge \neg C(x))$	$\begin{bmatrix} \neg r & A'_1 & \cdots & A'_m \\ \neg C & \neg C & \cdots & \neg C \end{bmatrix}$
$\hat{C} \sqsubseteq \breve{D}$, where $\hat{C} = \bigcap_{i=1}^{n} A_i$, $\breve{D} = \bigcup_{j=1}^{m} A'_j$, $A_i \in S_C$ (pure conjunction), $A'_j \in S_D$ (pure disjunction)	$A_1(x) \wedge \ldots \wedge A_n(x) \wedge \neg A'_1(x) \wedge \ldots \wedge \neg A'_m(x)$	$\begin{bmatrix} A_1 \\ \vdots \\ A_n \\ \neg A'_1 \\ \vdots \\ \neg A'_n \end{bmatrix}$

Table 2. Recursive sub-cases of existential and universal restrictions.

Axiom type	FOL Positive DNNF mapping	NNF Positive Matrix	Direct Matrix
A_i is an existencial restriction: $\ldots \sqcap \exists r.A \sqcap \ldots$, with $A \in S_C$ (pure conjunction)	$\ldots \wedge r(x,y) \wedge A(y) \wedge \ldots$	$\begin{bmatrix} \vdots \\ r(x,y) \\ A(y) \\ \vdots \end{bmatrix}$	$\begin{bmatrix} \vdots \\ r \\ A \\ \vdots \end{bmatrix}$
A'_j is an universal restriction: $\ldots \sqcup \forall r.A' \sqcup \ldots$, with $A' \in S_C$ (pure disjunction)	$\ldots \wedge r(x,y) \wedge \neg A'(y) \wedge \ldots$	$\begin{bmatrix} \vdots \\ r(x,y) \\ \neg A'(y) \\ \vdots \end{bmatrix}$	$\begin{bmatrix} \vdots \\ r \\ \neg A' \\ \vdots \end{bmatrix}$

3. An $\mathcal{ALC}$ Connection Calculus in Sequent Style

Definition 3 (Path, connection, unifier, substitution). A *path* is a set of literals from a matrix in which every clause (or column) contributes with one literal. A *connection* is a pair of complementary literals from different clauses, like $\{L_1^\sigma, \neg L_2^\sigma\}$, where $\sigma(L_1)$ (or $\sigma(\overline{L_2})$) is the most general unifier (mgu) between predicates L_1 and $\neg L_2$. σ is the set of

substitutions, which are mappings from variables to terms.

Definition 4 (Validity, active path, set of concepts). An $\mathcal{ALC}$ formula represented as a matrix is *valid* when every path contains a connection $\{L_1, \neg L_2\}$, provided that $\sigma(L_1) = \sigma(\overline{L_2})$. This is due to the fact that a connection represents the tautology $L_1^\sigma \vee \neg L_2^\sigma$ in DNF. As a result, the connection method aims at finding a connection in each path, together with a unifier for the whole matrix. During the proof, the current path is called *active path* and denoted by $\mathcal{B}$. The *set of concepts* τ of a variable or instance *x* during a proof is defined by $\tau(x) \stackrel{\text{def}}{=} \{C | C(x) \in \mathcal{B}\}$ [Schmidt & Tishkovsky 2007].

Definition 5 ($\mathcal{ALC}$ connection sequent calculus). Figure 1 brings the rules in sequent style of the $\mathcal{ALC}$ connection calculus, adapted from [Otten 2010].

$$Axiom\ (Ax)\ \frac{}{\{\}, M, Path}$$

$$Start\ Rule\ (St)\ \frac{C_2, M, \{\}}{\varepsilon, M, \varepsilon}$$

$$where\ M\ is\ the\ matrix\ KB \vDash \alpha,\ C_2\ is\ a\ copy\ of\ C_1 \in \alpha$$

$$Reduction\ Rule\ (Red)\ \frac{C^\sigma, M, Path \cup \{L_2\}}{C \cup \{L_1\}, M, Path \cup \{L_2\}}\ with\ \sigma(L_1) = \sigma(\overline{L_2})$$

$$Extension\ Rule\ (Ext)\ \frac{C_2^\sigma \backslash \{L_2^\sigma\}, M, Path \cup \{L_1\} \quad C^\sigma, M, Path}{C \cup \{L_1\}, M, Path}$$

$$with\ \ C_2\ a\ copy\ of\ C_1 \in M, L_2 \in C_2, \sigma(L_1) = \sigma(\overline{L_2}),$$

$$Copy\ Rule\ (Cop)\ \frac{C \cup \{L_1\}, M \cup \{C_2^\mu\}, Path \cup \{L_2\}}{C \cup \{L_1\}, M, Path \cup \{L_2\}}$$

$$with\ L_2 \in C_2, \mu \leftarrow \mu + 1, and\ \left(x_\mu^\sigma \notin N_O\ or\ \tau(x_\mu^\sigma) \nsubseteq \tau\left(x_{\mu-1}^\sigma\right)\right), \sigma(L_1) = \sigma(\overline{L_2})\ (blocking\ conditions)$$

Figure 1. The $\mathcal{ALC}$ connection calculus rules in sequent style (adapted from [Otten 2010]).

Blocking didn't occur in the original CM due to FOL semi-decidability, but it consists in a common practice in DL to guarantee termination. Here, to assure termination, we have to check if the set of concepts τ associated to the variable x_μ^σ (i.e., if the new x_μ was unified) of the new literal L_2^μ being created by the Cop rule is not contained in the set of concepts of the original *x* from $L_2(x)$ (in the rule, $\tau(x^\sigma)$) [Schmidt & Tishkovsky 2007]. Examples of the $\mathcal{ALC}$ CM calculus, as well as an algorithm of the system based on [Bibel 1987] can be found at [Freitas et al 2010].

In terms of complexity, the system is PSPACE in case of non-cyclical ontologies and EXPTIME for cyclical. Proofs of its completeness, soundness and termination are presented in [Freitas et al 2010].

Example 1 ($\mathcal{ALC}$ connection calculus).

$$\left.\begin{array}{l} \text{Animal} \sqcap \exists \text{hasPart.Bone} \sqsubseteq \text{Vertebrate} \\ \text{Bird} \sqsubseteq \text{Animal} \sqcap \exists \text{hasPart.Bone} \sqcap \exists \text{hasPart.Feather} \end{array}\right\} \vDash \text{Bird} \sqsubseteq \text{Vertebrate}$$

In FOL positive matricial clausal form, where the variables *y* and *t* were respectively skolemized by the function *f(x)* and the constant *c*, the formula is represented by

{{*Bird(x)* ,¬*Animal(x)*}, {*Bird(x)* ,¬*hasPart(x,f(x))*}, {*Bird(x)* ,¬*Bone(f(x))*}, {*Bird(x)* ,¬*hasPart(x,g(x))*}, {*Bird(x)* ,¬*Feather(g(x))*}, {*Animal(w), hasPart(w,z), Bone(z),* ¬*Vertebrate(w)*}, {¬*Bird(c)*}, {*Vertebrate(c)*)}}.

Figure 2 deploys the query proof. In the figure, literals of the active path are in boxes and arcs denote connections. For building a proof, we first choose a clause from the consequent (*Start rule*), say, the clause {¬*Bird(c)*} and a literal from it (¬*Bird(c)*).

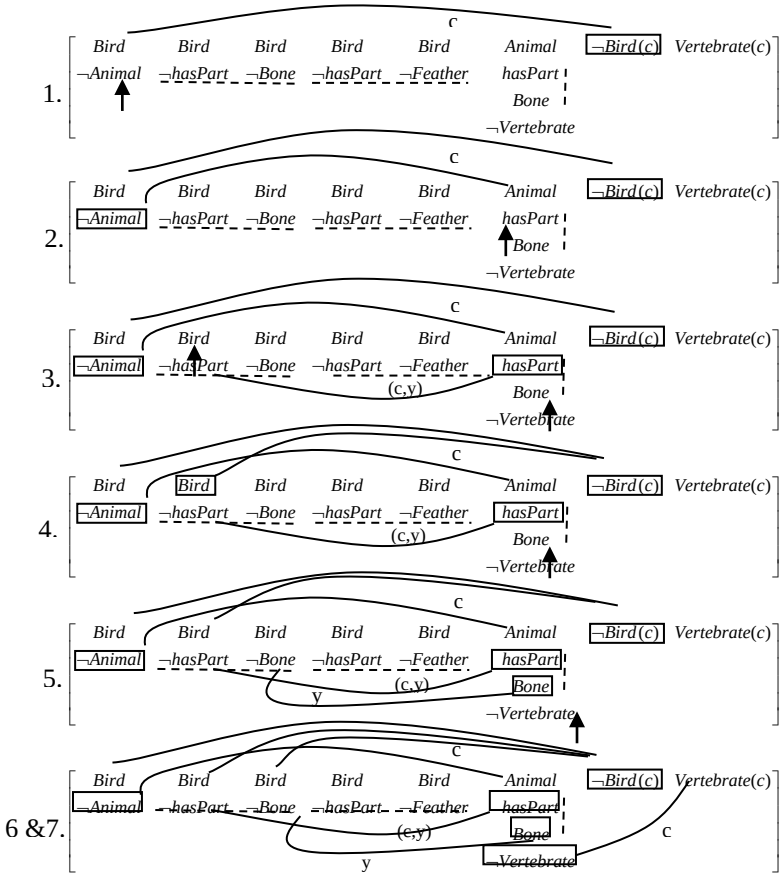


Figure 2. A connection proof example in matricial form.

Step 1 connects this clause with the first matrix clause. An instance or variable - representing a fictitious individual we are predicating about -, appears in each arc, for this connection, the instance *c*. The arrow points to literals to be checked in the clause (¬*Animal* in Step 1), that should be checked afterwards. After step 2, the connection {¬*Animal, Animal*} is not enough to prove all paths stemming from the other clause, the one with literal ¬*Animal*. In order to assure that, the remaining literals from that clause, *viz hasPart, Bone* and ¬*Vertebrate*, have still to be connected. Then, in step 3, when we connect *hasPart,* we are not talking about instance *c* any more, but about a relation between it and another variable or fictitious individual, say *y* (indicated by *(c,y)*).

Until that moment, we were only applying the *Extension rule.* However, in step 4, we use the *Reduction rule*, triggered by its two enabling conditions: (i) there is a connection for the current literal already in the proof; and (ii) unification can take place.

Unification would not be possible if we were referring to different individuals or skolemized functions (in $\mathcal{ALC}$, equality among individuals is not necessary).

A small note on unification is necessary here, because it brings a small trick to the calculus. Since horizontal dashlines represent universal restrictions ($\forall r.C$), the qualifier concept (C, represented as $\neg C$ in the matrix) correspond to a skolemized concept (say *C(f(c))*). Therefore, it can only be unified with variables, but not with concrete individuals or other skolemized qualifier concepts.

In case the system is able to summon the query, the processing finishes when all paths are exhausted and have their connections found. In case a proof cannot be entailed, the system would have tried all available options of connections, unifiers and clause copies, having backtracked to the available options in case of failure.

4. Conclusions and Future Work

I have formalized a connection method to take on the DL $\mathcal{ALC}$, by adapting the CM calculus formalized in sequent style from [Otten 2010] and including a new rule. I also introduced some notational improvements, the key one being the representation without variables. Of course, I plan to continue this work in many research directions, such as implementations, other DLs, Semantic Web, etc.

I intend to extend the work presented here to more complex description logic languages in a near future. Particularly, formalizations and implementations for the DLs $\mathcal{EL}$++, $\mathcal{SHIQ}$ and $\mathcal{SROIQ}$ will be practically useful for applications related to the Semantic Web and for some other biomedical applications that I am involved in.

Last but not least, lean implementations written in Prolog, in the flavor of leanCop [Otten & Bibel 2003], that demand small memory space, can serve applications that are constrained in memory, such as stream reasoning in mobile applications, for instance. They are also in my research agenda.

References

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (Eds.): The Description Logic Handbook. Cambridge University Press, 2003.

Bibel, W. Automated theorem proving. Vieweg Verlag, Wiesbaden, 1987.

Freitas, F. A Connection Method for Reasoning with the Description Logic $\mathcal{ALC}$. Technical report. 2010. www.cin.ufpe.br/~fred/CM-ALCTechRep.doc

Otten, J.Restricting backtracking in connection calculi.AI Comm,23(2-3):159-182 2010.

Otten, J., Bibel, W. *leanCoP: Lean Connection-Based Theorem Proving*. Journal of Symbolic Computation, Volume 36, pages 139-161. Elsevier Science, 2003.

Schmidt, R., Tishkovsky, D. Analysis of Blocking Mechanisms for Description Logics. In Proceedings of the Workshop on Automated Reasoning, 2007.

Collaborative Construction of Visual Domain Ontologies Using Metadata Based on Foundational Ontologies

Gabriel M. Torres, Alexandre Lorenzatti, Vitor Rey, Rafael P. da Rocha, Mara Abel

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

(gmtorres, alorenzatti, vfrey,marabel)@inf.ufrgs.br, rafael.rocha@ufrgs.br

Abstract. *Domain ontologies are widely used to explicit declarative knowledge. However, it is a difficult task to obtain an explicit and shared vocabulary that can be used in computer systems. Besides that, many domains require not only textual data but also visual data to express the meaning of the concepts. Some ontology editors have been developed to support collaboration on the ontology development process, but none of them have support to visual domains or ontological foundation, which is important to build efficient ontologies with semantic validation. In this paper, we introduce an approach to support collaborative construction and evolution of visual domain ontologies using metadata ontologies based on a foundational ontology.*

1. Introduction

In the Knowledge Engineering process, we are usually concerned on collecting knowledge shared by a specific community, storing it as a formal ontology and using this ontology as a reusable artifact for various purposes. As we know so far, an ontology is defined as a formal specification of a *shared* conceptualization (Borst, 1997). Knowledge domains are not static: they evolve when new elements become part of the domain or when elements become obsolete (De León, 2009). These changes need to be adapted to the domain model, updating the ontology by adding or removing components. Therefore, collaboration has become an important part of the ontology development process, helping in making explicit the concept behind vocabulary and evolving the vocabulary to its new meanings.

Several classic ontology editors have been developed over time. Some interesting recent approaches focus on collaboration aspects of ontology development, like NeON Project (Haase et al., 2008), OntoEdit (Sure et al., 2002) and Collaborative Protégé (Tudorache, Noy, & Musen, 2008). Collaborative Protégé has introduced a metadata ontology to support this collaboration, allowing the specification of changes and annotations on the domain ontology. However, we think that the formal language oriented-interface is hard for the domain specialist to deal with, because he/she is not usually familiar with ontology formalization.

Many information domains like Medicine and Geology need visual information to express knowledge. Our tool allows the user to build a conceptual model of visual knowledge, getting advantage of visual representations like icons and images to help in expressing the full meaning of the concepts. On the other hand, in order to capture the correct meaning of a concept, it is necessary to let the user express his/her understanding about that concept through the use of properties that have concrete meaning to him/her. This is the role of foundational ontologies: express the inherent properties that provide identity to the objects in some world. Our proposal consists of developing a collaborative web environment for ontology construction with two main contributions: support to visual domains and support to ontological foundation.

The collaboration is based on metadata information about the ontology components, so that the users can express their understanding about the meaning of the concepts without requiring any formal representation language, but giving mechanisms to manipulate visual information and to express rich semantic models. Metadata information provides the necessary vocabulary and artifacts that can offer the basis for the development of applications for collaborative ontology construction. The metadata and the domain ontology data are generically stored in a database as triples.

This paper is organized as follows: In Section 2, we explain why and how visual domains are important on ontology development. In Section 3, we present the foundational ontology that provides ontological foundation to our metadata models. Section 4 introduces the metadata ontologies that are the main contribution of this work. In Section 6, we conclude and anticipate some future work.

2. Visual Domains

Some information domains require visual knowledge as a crucial part of the problem solving process like Medicine or Geology and most of Natural Sciences. The interpretation process occurs through a visual pattern matching against the domain, capturing the objects that can support the inference path. Some of these visual objects have even barely translation to a propositional description. Therefore, the ontology construction in visual domains requires more than symbolic descriptions to explain the concepts. Besides that, many ontology developers find it easier to provide descriptions of their domain concepts and properties using visual representations rather than only formal descriptions (or pure natural language descriptions).

According to Lorenzatti (2011), a concept can be represented in two different ways: through a symbol from a language or through a pictorial representation (an imagistic representation, like an image or icon). Icons are similar to what they represent, so, their meaning is captured through the same process of perception used to recognize the represented object or event. In other words, its meaning can be understood from the observation of the representation. Images are photographs of concepts, intended to provide examples of instances of a concept, trying to transmit its meaning. Therefore, the concepts can be associated either to a symbol or to a pictorial representation, like an image, a draw, an icon, a chart, etc. An icon can also be associated to each property value. For example, the property *Roundness* can have an associated icon to each of its values: low-rounded, rounded, high-rounded.

3. Foundational Ontology

The aim of a foundational ontology is establishing a basis to obtain coherence in the negotiations of meaning during the collaboration process of individuals to build a conceptual model. Recently, a *Unified Foundational Ontology (UFO)* was proposed (Guizzardi 2005), defining categories that provide ontological foundation in the construction of conceptual models. The UFO is divided in three fragments called UFO-A, UFO-B and UFO-C. We are interested in the theoretical framework of meta-properties and meta-types proposed by Nicola Guarino (Guarino 1995) and Giancarlo Guizzardi, mainly focused in the UFO-A, which is the core of the foundational ontology, consisting of a stable theory that introduces structuring concepts to offer more semantics to conceptual modeling languages. Therefore, we will mention here the notions of *rigid sortals*, *properties*, *quality domains*, *partonomic relations* and *hierarchical relations*. For instance, a rigid sortal is a concept whose definition requires that their instances cannot stop being an instance of this concept in any possible world. This means that if the *essential properties* chosen to define the concept cease to be recognized in the way they were defined, the instance will cease to exist to because it loses its identity criterion. A person is a rigid sortal while a student is not, since there are instances of it that can stop being a student without losing its identity. These are important constructs for ontological models, since they allow producing trustful mappings among different domain ontologies that support interoperability.

The current ontology development tools don't implement this rich formal semantic representation because they are based only in the five basic ontology constructs (concept, property, property value, relation, axiom) that don´t express the differences of objects in reality according to human discrimination. The unified foundational ontology extended these basic primitives, creating several additional constructs that helps in establishing the taxonomic classification and the relationships among concepts. Therefore, ambiguity is reduced and the expressivity of the model is increased.

4. Metadata Ontologies for Collaboration

In this paper, we introduce an upper-level domain independent metadata to specify the structure of the domain ontology components and collaboration events. Using this metadata, the community of users can define concepts, attributes and domain values, making also explicit the intended meaning through the use of primitives of a foundational ontology (assigning values to meta-properties and meta-types of concepts), visual icons and illustrative images. The conflicts about selection of names, attributes, icons and images are solved by the proposition of changes in the models. The changes are justified by the ontological definition and are stored for further reference.

We introduce two metadata ontologies: the Representation Ontology (R.O.), which defines primitives for representing the domain ontology, and the Collaboration Ontology (C.O.), which defines primitives for representing the collaboration events. The domain ontology components are defined as instances of the R.O. concepts and the changes made over the domain ontology are defined as instances of the C.O. concepts. These models are the basis of our environment, structuring the meta-level data and helping the application to deal with the abstract representations of the domain ontology and to track changes involving symbolic or visual representations and foundational artifacts.

4.1. Representation Ontology (R.O.)

When dealing with ontologies, we are commonly focused on its main components: concept, property, property value, relation and axiom. To add more semantics, this meta ontology extends some of the main ontology components by specializing them based on visual and foundational aspects.

In order to provide visual support, we used the concept [Image], which is specialized in two sub-concepts: [Photography] (for representing photos of concept instances) and [Icon] (for representing symbolic pictorial icons). The R.O. contains some relations that link a [OntologyConcept] or [OntologyPropertyValue] to one [Icon] (*hasIcon*) and a [Photography] to a [OntologyConcept] (*photographyOf*). To provide ontological foundation to the model, we have specialized the R.O. concepts using some of the foundation constructs proposed in the UFO-A foundational ontology, enriching the semantics of the model without adding significant complexity. The [OntologyConcept] was specialized to represent Substantial Universals and its subclasses: Sortal, Kind, Mixin, RigidSortal, etc. The [OntologyProperty] concept was specialized to represent the distinct types of property: DataTypeProperty (for properties that point to primitives like *string*, *int*, *datetime*, etc.) and QualityUniversalProperty (for properties that have one or more pre-defined values, like color, age interval, etc.). The [OntologyPropertyValue] concept was specialized in Quale. The [OntologyRelation] concept was specialized to represent the different types of relations, allowing the representation of partonomic relations (ExtensionOf, MemberOf, PartOf, SetOf, SubQuantityOf and SubsetOf) and the hierarchical subsumption relation (SubclassOf). These constructs, when used correctly, impose semantic restrictions to the model, which can be analyzed by the knowledge engineer to help the users to detect semantic failures and representational misuses on the domain ontology.

4.2. Collaboration Ontology (C.O.)

The Collaboration Ontology (C.O.) defines which collaboration activities can be done on the domain ontology. In a simplified way, the C.O. instances are the changes related to what has been represented by the R.O. ontology, which is the domain ontology. The collaboration process is focused on the proposal and storage of changes made over concepts, properties or relations and also on annotations that can be attached to any domain ontology component. The specialists can make directly changes or annotations in the domain model by adding, changing or removing ontology components. Collaboratively, they can see each other change history and discuss about it, possibly making new modifications until a consolidated domain model is obtained.

The C.O. concepts define the set of possible ontology changes that can be done in the system. A change event has some properties that give meaning to it: *domainComponent* relates the change to one domain ontology component; *author* stores who made the change; *date* stores the date and time when the changed occurred, to help in tracking the evolution of ontology; *value* stores the new value of the change. The C.O. describes not only common changes (ConceptCreated, PropertyRemoved) but also visual changes (ConceptIconChange, ConceptPhotographyChange) and foundational changes (MemberOfRelationCreated, SubsetOfRelationCreated, QualeCreated, QualeIconChanged, etc.). Therefore, we can change the semantics of the concepts,

adding more information to the domain ontology model than the current approaches offer. For example, if the user once created a concept by instantiating the R.O. concept [OntologyConcept], and now he/she wants to change its stereotype because it is, in fact, a RigidSortal, it can be done by creating an instance of the C.O. [ConceptTypeChange] concept and setting its *value* property to "RigidSortal". The concept type can be changed unlimited times by the community, to reach the correct consensual semantics.

The C.O. also allows the collaboration on visual domains. Icons are an unique alternative representation of the concept, based in the visual perception, that helps in avoiding the excessive use of propositional interfaces in ontology-based systems. When a user changes a concept icon, an instance of [ConceptIconChange] is created and associated both to the concept and to the icon image uploaded. Using the same procedure, a concept can be associated to one or more photographs by creating instances of the C.O. concept [ConceptPhotographyCreated]. If a user deletes a photograph, an instance of [ConceptPhotographyRemoved] is created. A property value can also have an associated icon by creating an instance of [QualeIconChanged]. Further, the users can make comments specifically about the icons or photos and change these artifacts later on until they reach the correct consensual visual representation. In Lorenzatti (2011), a library of icons related to the Sedimentary Geology domain was developed with cognitive analysis support. We are currently using this library to validate our environment. An example of the metadata ontology interaction and the collaboration history generated when changing visual and foundational aspects of the domain ontology is shown in Figure 1.

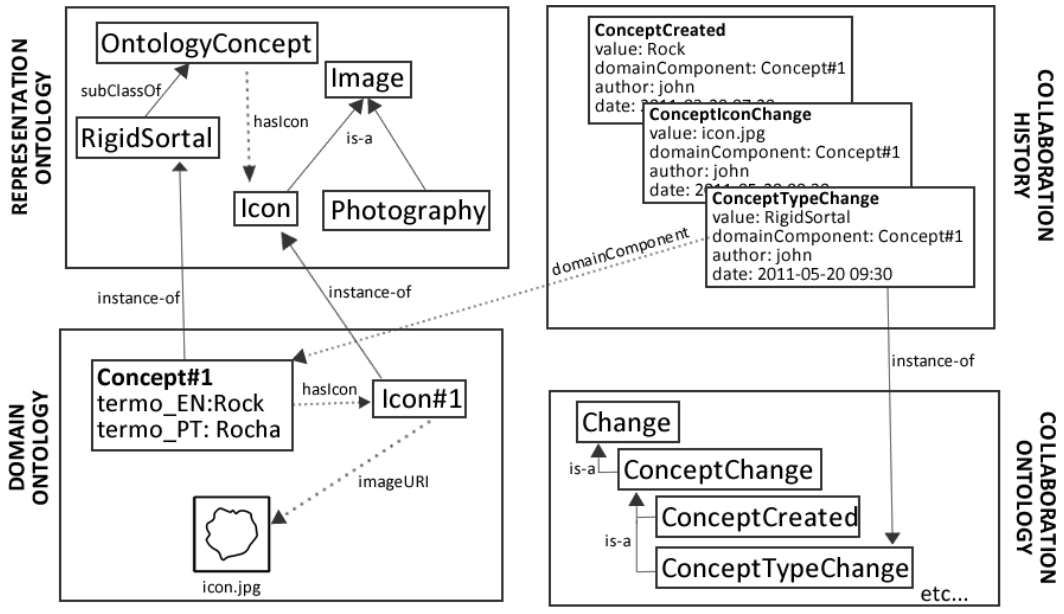


Figure 1 Collaboration structure with visual and foundation constructs

6. Conclusions and Future Work

Domain ontologies are explicit conceptual models of a shared knowledge, focused on a specific domain of common interest. Some domains need more than textual representations for concepts of reality, requiring visual representations as well. The foundational concepts introduced in the UFO-A foundational ontology helps the

construction of semantic models by helping to establish the essential properties that express the identity of the concepts. Besides that, the foundational data helps the knowledge engineer to understand the domain and to interact with the specialists. We consider that supporting the user to recognize these properties would help in doing ontological choices in associating the concepts with the representation constructs. This would lead to the development of better quality domain ontologies, in terms of lucidity and laconicity, avoiding ambiguity and redundancy. In order to provide the adequate support, we develop a framework based on metadata ontologies that allow the domain experts in developing and evolving domain ontologies using textual and visual representations and ontological foundational data.

The metadata ontologies introduced in this paper and its usage are focused on providing a basis for the collaborative construction of rich language-independent domain ontologies. The generated collaboration events form an important collaboration and evolution history that can be used to analyze critical points of the ontology or to track changes. Our approach is experimentally available as a web-based environment for the collaborative construction of the Sedimentary Geology domain ontology at the address *http://obaita.inf.ufrgs.br/*, which is currently under constant development.

7. Acknowledgements

This work had the scholarships supported by CAPES and the financial support of CTPETRO Government program and ENDEEPER Co.

8. References

BORST, W. N. **Construction of Engineering Ontologies for Knowledge Sharing and Reuse.** Twente: University of Twente, 1997. Doctoral thesis.

DE LEÓN, P.. **Ontology metadata management in distributed environment**. Madrid:Universidad Politécnica de Madrid, 2009. Doctoral thesis.

GUARINO, N. Formal ontology, conceptual analysis and knowledge representation. **Int. Journal of Human Computer Studies**, v.43, n.5, p. 625–640,1995.

GUIZZARDI, G. Ontological Foundations for Structural Conceptual Models. **Enschede**, The Netherlands: Universal Press, v.5, n.74. 2005. 410 p. (CTIT PhD Thesis Series).

HAASE, P. et al. The Neon Ontology Engineering Toolkit. In International World Wide Web Conference, 17, Beijing, China, April, 2008, **Proceedings**…, 2008 (Developers Track.)

LORENZATTI, A.;ABEL, M.; FIORINI, S.;BERNARDES, A.;DOS SANTOS SCHERER, C..Ontological Primitives for Visual Knowledge. In Brazilian Conference on Advances in artificial intelligence,São Bernardo do Campo, 2010. **Proceedings**… Berlin / Heidelberg:Springer, , 2011. (Lectures Notes in Artificial Intelligence v.6404,)

SURE, Y. et ali.OntoEdit: Collaborative ontology development for the semantic web. In The International Semantic Web Conference,1, Sardinia, Italy, June 2002. **Proceedings**…. 2002. p.221-235.

TUDORACHE, T.; NOY, N. F.; MUSEN, M. A.Collaborative Protege: Enabling community-based authoring of ontologies. In International Semantic Web Conference, 7, Karlsrhue, Germany, 2008. **Proceedings**… 2008. (Posters & Demos).

The Limits of Using FrameNet Frames to Build a Legal Ontology

Anderson Bertoldi, Rove Luiza de Oliveira Chishman

Applied Linguistics Graduate Program
Universidade do Vale do Rio dos Sinos (UNISINOS) – São Leopoldo, RS – Brazil
andersonbertoldi@yahoo.com, rove@unisinos.br

Abstract. *FrameNet frames have been used to develop lexical databases and annotated corpora for different languages. This paper analyses the use of FrameNet frames to build a legal ontology for the Brazilian Law. In order to discuss the problems of such approach to ontology development, the lexical units evoking the* `Criminal_process` *frame were contrasted in English and Portuguese. Frame divergence between languages has consequences not only for legal ontology development but also for the development of legal lexical resources, such as lexical databases and corpora annotation.*

1. Introduction

Semantic Web technology for the legal domain has been an important topic in the last years. Semantic Web technologies involve both applications in corporate settings, such as knowledge management and intranet systems, and public information retrieval on internet (Benjamins at al., 2005). Semantic lexicons and legal ontologies have been developed to facilitate the access to legal information.

Lexicons and ontologies sometimes are considered as a similar resource. The parallel between word sense and ontological categories in one hand and lexical relations and ontological relations on the other hand suggests the similarity of these two resources (Hirst, 2003). Nevertheless, ontologies and lexicons are different resources. In the words of Gruber (1993, p.199), an ontology is an explicit specification of a conceptualization. While lexicons represent words senses in a natural language, ontologies are, by definition, an engineering artifact that represent the knowledge of a particular area in a formal language.

This paper analyses the use of FrameNet frames to build a legal ontology for the Brazilian Law[1]. It is a first attempt to construct a legal ontology lexically oriented. In this paper, the lexical units evoking the `Criminal_process` frame are contrasted in English and Portuguese. The aim of this contrastive study is discussing the conceptual structure evoked by lexical units and how this information is particular to each country. In social-oriented areas, such as Law, concepts may not be shared among countries. Considering countries that share the same language, a legal concept may evoke different

[1] The work presented here was developed in the scope of the of the project *Semantic Technologies and Legal Information Retrieval Systems*, a project supported by CAPES and CNJ (Conselho Nacional de Justiça) and coordinated by Professor Dr. Rove Luiza de Oliveira Chishman.

conceptual structures, because countries do not share the same set of laws and regulations.

FrameNet frames have been used to develop lexical databases and annotated corpora for different languages. Semantic frames are considered conceptual structures independent of language (Boas, 2005; Padó, 2007). As conceptual structures independent of language, semantic frames would have the characteristic of being universal. This is the principle that enables the transfer of semantic annotation from corpus of one language to another language (Padó, 2007) and the automatic development of lexicons expanding FrameNet frames to other languages other than English (Padó e Lapata, 2005).

This paper demonstrates that FrameNet frames are not always language-independent conceptual structures. In order to discuss the problems of using FrameNet frames to build legal ontologies, this paper is structured in seven sections. Section 2 presents legal ontologies and lexicons. Section 3 presents FrameNet methodology for frame creation. Section 4 presents the methodology for FrameNet creation other than English. Section 5 presents the FrameNet `Criminal_process` frame. Section 6 presents the mismatches between legal knowledge in USA and Brazil. Section 7 presents the conclusions of this work.

2. Legal Ontologies and Lexicons

Terminological lexicons and legal ontologies have been proposed for legal information retrieval purposes. *Core Legal Ontology* (CLO) (Gangemi et al., 2005) is an ontology developed by the Institute for Theory and Techniques for Legal Information (ITTIG-CNR). This ontology is used to structure legal concepts from the terminological lexicon *JurWordNet* (Gangemi et al., 2005). *LRI-Core* (Breuker et al., 2005) is a legal ontology developed by the Leibniz Center for Law, in the scope of the European project e-COURT (Breuker et al., 2005). The main purpose of LRI-Core is to support knowledge acquisition to legal domain ontologies and allow automatic indexing of legal documents.

Terminological wordnets like *Jur-WordNet* (Sagri et al., 2004) aim to improve legal information retrieval by connecting terms through semantic relations, mainly synonymy. LOIS (Lexical Ontologies for Legal Information Sharing) (Curtoni et al., 2005) was an investigation project supported by European Commission within the e-Content program. The aim of LOIS was to build a European legal wordnet for legal information retrieval. The semantic relations connect terms in different languages. The LOIS architecture was based on another European project, the *EuroWordNet* (Vossen, 1998). In LOIS the different language databases were connected through an interlingual index. This work differs from lexicons and ontologies presented in this section because it aims at using lexical databases to develop legal ontologies.

3. FrameNet

FrameNet is a lexical database that describes word meaning according to the principles of Frame Semantics. In FrameNet lexical items are conceived as lexical units. A lexical unit is the combination of a word form with a meaning. Every new meaning of a word represents a new lexical unit. Therefore, it is the lexical unit that evokes the frame, not

the word. According to Fillmore and Baker (2010), the method of lexical analysis in FrameNet follows five steps: (1) **Characterizing the frames**, (2) **Describing and naming frame elements**, (3) **Selecting lexical units**, (4) **Creating manual annotations of sample sentences** and (5) **Automatically generating lexical entries**.

4. Methodology for FrameNet Creation

In order to discuss the use of FrameNet frames for legal ontology development, it is necessary to present some points related to FrameNet and multilinguality. FrameNet for languages other than English has been created using the expansion methodology. Expansion methodology assumes that semantic frames stay the same and only the linguistic information is substituted to create new FrameNets. This is the methodology adopted by Spanish FrameNet (Subirats, 2009) and Japanese FrameNet (Ohara, 2009). According to Lönneker-Rodman (2007, p.5), expansion methodology risk to "(…) neglecting language-specific differences in lexicalization". Lönneker-Rodman (2007) presents four types of mismatches between frames in FrameNet construction: (1) **Semantic Frame**, (2) **Frame Elements**, (3) **Semantic Type and Frame Element Coreness** and (4) **Frame Relations**. The criteria for new frame creation presented in Lönneker-Rodman (2007) show that lexical changes between two languages will change the conceptual structure, in other words, lexical changes may affect the structure of a semantic frame. This work analyses how different the conceptual structure may be in a social-oriented field like Law.

5. `Criminal_process` Frame

In the FrameNet terminology, `Criminal_process` frame is a non-lexical frame. The function of non-lexical frames is to connect semantically related frames. Non-lexical frames do not present frame-evoking lexical units. They represent complex events divided in more specific frames. `Criminal_process` frame describes the different steps of a criminal process according to the American legal system. In case of complex frames, like `Criminal_process`, each sequence of events or states is described as a single frame, related to the complex frame through *Subframe* relations and to the other subframes through *Precedes* relation. `Criminal_process` frame is divided in four subframes temporally succeeded: `Arrest`, `Arraignment`, `Trial`, and `Sentencing`. `Arraignment` frame is divided in three subframes: `Notification_of_charges`, `Entering_a_plea`, and `Bail_decision`. `Trial` frame also presents three subframes: `Court_examination`, `Jury_deliberation` and `Verdict`. There is still the frame `Try_defendant`. In FrameNet terminology, the `Trial` and `Try_defendant` frames are in a *Perspective* relation. That relation describes frames that are similar and represent two sides of the same event. Therefore, the `Trial` frame describes the organization of the trial, while the `Try_defendant` frame describes the event of trying a defendant.

6. Mismatches between Legal Knowledge in USA and Brazil

In order to contrast the legal knowledge described by `Criminal_process` frame to Brazilian legal system, firstly it was necessary to create manually frames to represent the legal information about a criminal process according to the Brazilian legal system.

Criminal_process frame was contrasted with a Brazilian criminal process frame considering three levels of linguistic analysis: lexical units, frames, and frame elements. In the contrastive study, it is possible to perceive that semantic frames present different levels of equivalence. Some frames in FrameNet found an equivalent frame in Brazilian legal system, with lexical units presenting equivalents in Portuguese, correspondence between the American and Brazilian legal knowledge, and the same frame elements for both frames in English and Portuguese. Other FrameNet frames found equivalence only between lexical units in English and Portuguese, that is, the legal event represented in FrameNet frame did not exist in Brazilian Legal system.

The FrameNet frame Try_defendant represents a legal event in which a defendant is tried by a jury or a judge in a court (FrameNet definition). The core frame elements for Try_defendant frame are: CHARGES, DEFENDANT, GOVERNING_AUTHORITY, JUDGE, and JURY. The lexical unit that evokes this frame is *to try* that has as an equivalent in Portuguese the lexical unit *julgar*. The legal event represented by Try_defendant frame is comparable to the legal event of trying a defendant in Brazil. The lexical unit *julgar* in Portuguese evokes a legal knowledge comparable to the legal knowledge evoked by *to try*. It is possible to say that Try_defendant is equivalent to the Brazilian legal frame Julgar_acusado.

Other frames, like Notification_of_charges, present only lexical unit equivalence. Notification_of_charges represents a legal event in which the judge informs the accused of the charges against him/her (FrameNet definition). The core frame elements for Notification_of_charges are: ACCUSED, ARRAIGN_AUTHORITY, and CHARGES. The lexical units that evoke Notification_of_charges frame are: *to accuse*, *charge*, *to charge*, *to indict*, and *indictment*. These lexical units present equivalent in Portuguese: *to accuse/acusar*, *charge/acusação*, *to charge/acusar*, *to indict/pronunciar*, and *indictment/pronúncia*, but the legal knowledge evoked by these lexical units in English is not the same legal knowledge evoked by their equivalents in Portuguese.

There are still frames that do not present equivalence of any type. This is the case of Arraignment frame. Arraignment frame describes a legal event that is typical of the American system which is based on Common Law. Even the frame-evoking lexical units do not find an equivalent in Portuguese: *arraign* and *arraignment*.

7. Conclusion: Consequences of Frame Mismatches for Ontology Development

The conception of semantic frame as a language-independent conceptual level motivates proposals to use semantic frames as interlingual representation for multilingual databases (Boas, 2005). Other proposal for semantic frames is automatic generation of frame-based lexical databases (Padó, 2007). Semantic frames are still used for corpus annotation in languages other than English (Burchardt et al., 2009). Considering the use of FrameNet frames for lexical resource development, it is important to ask whether FrameNet could be a starting point even for ontology development. The assumption that semantic frames are conceptual structures language-independent could enable the use of semantic frames as ontological categories.

The contrastive study of `Criminal_process` frame brought important evidences against the use of FrameNet frames for legal ontology development. Semantic frames may represent a more language-independent conceptual level, but being language-independent does not mean being social or cultural-independent. Law is a social-oriented area, which means that laws are not equal for all countries. The frame evoked by a lexical unit will reflect the legal knowledge of a regulation from a specific country. In countries that speak the same language, a lexical unit may evoke a different legal knowledge in both countries.

The levels of mismatches between frames in different languages show that semantic frames cannot be considered a conceptual level language and cultural-independent. An example of semantic frame that does not exist in Brazilian legal system is the `Arraignment` frame. The arraignment is a hearing in which a suspect is asked to entering a plea. Therefore, the lexical units that evoke `Arraignment` frame in English do no find a translation equivalent in Portuguese. In other cases, some parts of a frame do not find a correspondence in Brazilian criminal process. This is the case of `Trial` frame. `Trial` frame is divided in three subframes: `Court_examination`, `Jury_deliberation`, and `Verdict`. According to the Brazilian legal system, the `Jury_deliberation` frame is not a step in a trial.

FrameNet semantic frames are not a good source of concepts for a legal ontology. Legal frames are social and culturally oriented. Using FrameNet frames would mean adopting a conception of legal system based in the United States legal organization. Different from other works in development of lexical resources, the social character of legal frames make them being specific for each society. Using FrameNet frames to develop an ontology to cover the concepts of the Brazilian legal system would entail the problem of adapting frames created for American legal system. A possible solution until this moment would be the manual modeling of a Brazilian legal ontology.

References

Benjamins, V. R., Casanovas, P., Breuker, J. and Gangemi, A. (2005). Law and the Semantic Web, an Introduction. *In*: Benjamins, V. R., Casanovas, P., Breuker, J. and Gangemi, A. (Eds.). *Law and the Semantic Web*: Legal ontologies, methodologies, legal information retrieval, and applications, LNAI (3369). Berlin/Heidelberg: Springer-Verlag, p.1-17.

Boas, H. C. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography.* Vol. 18, No.4, p. 445-478.

Breuker, J. A., Valente, A., and Winkels, R. (2005). Use and Reuse of Ontologies in Knowledge Engineering and Information Management. *In*: Benjamins, V. R., Casanovas, P., Breuker, J. and Gangemi, A. (Eds.). *Law and the Semantic Web:* legal ontologies, methodologies, legal information retrieval, and applications. Lecture Notes on Artificial Intelligence 3369. Springer-Verlag, Berlin/Heidelberg, p.36-64.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2009). Using FrameNet for the semantic analysis of German: annotation, representation, and automation. In: Boas, H. (Ed.) *Multilingual FrameNets* - Methods and Applications. Berlin/New York: Mouton de Gruyter, p.209-244.

Curtoni, P., Dini, L., Di Tomaso, V., Mommers, L., Peters, W., Quaresma, P., Schweighofer, E. and Tiscornia, D. (2005). Semantic access to multilingual legal information. *In*: Schweighofer, E. (Ed.). Proceedings of EU Info Workshop "Free EU Information on the Web: the future beyond the new EUR-LEX" of JURIX 2005 – The 18 th Annual Conference on Legal Knowledge and Information Systems. Brussels: Vrije Universiteit, pp.1-11. Available at: www.di.uevora.pt/~pq/papers/eu-ws-lois.pdf.

Fillmore, C.J., and Baker, C. (2010). A Frames Approach to Semantic Analysis. *In*: B. Heine, B. and Narrog, H. *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, p. 313-339.

Gangemi, A., Sagri, M. T., and Tiscornia, D. (2005). A Constructive Framework for Legal Ontologies. *In*: Benjamins, V. R., Casanovas, P., Breuker, J. and Gangemi, A. (Eds.) *Law and the Semantic Web:* legal ontologies, methodologies, legal information retrieval, and applications. Lecture Notes on Artificial Intelligence 3369. Springer-Verlag, Berlin/Heidelberg, p.97-124.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*. v.5, n.2, p.199-220.

Hirst, G. (2004). Ontology and the lexicon. *In*: Staab, S.; Studer, R. (Eds). *Handbook on Ontologies*. Springer: Berlin, 2004, p. 209-229.

Lönneker-Rodman, B. *Multilinguality and FramNet*. Technical Report. TR-07-001. Berkeley: ICSI, 2007.

Ohara, K. H. (2009). Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru. In: Boas, H. C. (Ed.) *Multilingual FrameNets in computational lexicography:* Methods and applications. Berlin/New York: Mouton de Gruyter, p.163-182.

Padó, S. (2007). *Cross-lingual Annotation Projection Models for Role-Semantic Information*. PhD Thesis. Saarbrücken: Universität des Saarlandes.

Padó, S. and Lapata, M. (2005). Cross-lingual projection of role-semantic information. In: *Proceedings of HLT/EMNLP-05*, Vancouver: Association for Computational Linguistics, 2005, p.859-866.

Sagri, M. T., Tiscornia, D. and Bertagna, F. (2003). Jur-WorNet. In: P. Sojka et al. (Eds.) Second International Wordnet Conference - GWC 2004. Brno: Masaryk University, p.305-310.

Subirats, C. (2009). Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In: Boas, H. C. (Ed.) *Multilingual FrameNets in computational lexicography:* Methods and applications. Berlin/New York: Mouton de Gruyter, p.136-162.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*. Vol. 32, Nº. 2-3, p.73-89.

Tesauro conceitual e ontologia de fundamentação: análise de elementos similares em seus modelos de representação de domínios

Jackson da Silva Medeiros[1], Maria Luiza de Almeida Campos[2]

[1]Programa de Pós-Graduação em Comunicação e Informação – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[2]Programa de Pós-Graduação em Ciência da Informação – Universidade Federal Fluminense (UFF)
Niterói – RJ – Brazil

jacksonmedeiros@yahoo.com.br, maria.almeida@pq.cnpq.br

Abstract. *Através do estudo de modelos conceituais de ferramentas utilizadas na Ciência da Informação e na Ciência da Computação, este trabalho objetiva analisar, de forma comparativa, os elementos semelhantes entre os modelos conceituais utilizados na construção de tesauros conceituais e de ontologias de fundamentação, a partir de um modelo de agregação proposto na literatura de Ciência da Informação.*

Resumo. *From the study of conceptual models from tools used in Library and Information Science and Computer Science, this study aimed to analyzing in a comparative way, the similar elements between the conceptual models used in the conceptual thesauri and foundational ontologies, from an aggregation model proposed in Library and Information Science literature.*

1. Introdução

A construção de modelos conceituais está diretamente ligada à representação do conhecimento e estes devem ser capazes de representar um contexto, sendo construídos a partir de processos que evitem qualquer tipo de ambiguidade, ressaltando objetos relevantes ao domínio, bem como seus relacionamentos e atributos. Na Ciência da Informação, a representação de domínios do conhecimento é responsável pela organização e recuperação de conhecimento registrado. Assim, com a necessidade cada vez maior de recuperação de informações de forma consistente, deve ser levado em conta os processos teóricos e metodológicos que permitem desenvolver modelos capazes de organizar e representar conhecimento.

É importante que a construção de sistemas que permitem recuperação da informação seja baseada no conhecimento existente, permitindo que o conhecimento sobre o mundo permita a construção de modelos sobre uma realidade. Esses modelos são representações parciais de determinado mundo, onde é possível representar a existência de objetos e as relações entre eles, gerando estruturas processáveis por máquina, quando se trata questões computacionais, e permitindo a construção de linguagens documentárias, quando se trabalha com questões na Ciência da Informação.

A Ciência da Informação vem trabalhando questões teóricas e metodológicas capazes de fornecer bases para a construção de modelos conceituais e, consequentemente, sistemas de organização e representação do conhecimento – como os tesauros conceituais –, permitindo que modelos sejam organizados a partir de conceitos e categorias, garantindo a durabilidade do sistema ao comportar sua atualização. Isto está pautado em questões que tratam o conceito, a partir dos estudos de Ingetraut Dahlberg (1978a, 1978b), além de trabalhar com a categorização dos mesmos, baseados na Teoria da Classificação Facetada, de Shiyali Ramamrita Ranganathan (1967).

O tesauro conceitual é formado por uma parte alfabética, onde os termos são apresentados na forma alfabética com as especificações das relações existentes entre eles, e uma parte sistemática, onde os conceitos se apresentam no modelo conceitual do instrumento. Além disso, esses tesauros se preocupam com o conteúdo conceitual dos termos, o que destaca a importância das definições de cada conceito.

No que tange a Ciência da Computação, seu foco representacional é a possibilidade de realizar comunicação entre sistemas. Nos últimos anos, porém, parecem ter havido percepções que possibilitam a modelagem de parte de um domínio tendo como base teorias independentes de domínio, como as ontologias de fundamentação (GUIZZARDI, 2005), tornando-se importante para a elaboração de modelos conceituais, aplicando teorias filosóficas e cognitivas neste processo e fornecendo princípios ontológicos para classificação de conceitos.

Uma ontologia de fundamentação apresenta princípios que estão concernentes com uma ontologia formal. Esses princípios, independentes de um domínio, permitem a elaboração de modelos para a representação de diversos contextos de representação, sendo altamente reutilizáveis. É também caracterizada por ser filosoficamente bem fundamentada, permitindo a explicitação de uma visão da realidade, ou seja, do acordo ontológico estabelecido, com determinação de regras de restrição, bem como conceitos, categorias e metapropriedades.

Este trabalho procura analisar de forma comparativa os modelos conceituais em que tanto tesauros conceituais quanto ontologias de fundamentação estão pautadas, de forma a promover a verificação das semelhanças encontradas em seus modelos de representação de domínios, tomando por base princípios estabelecidos por Campos (2004) para a modelagem conceitual de domínios.

2. Metodologia

A comparação entre os modelos conceituais de tesauros conceituais e ontologias de fundamentação foi realizada a partir de um procedimento sistemático e organizado, possibilitando estabelecer relações de semelhança entre objetos, a fim de concluir algo (COLINO, 2002). Privilegia-se, neste trabalho, o processo (comparativo) mais do que o modelo em si, evidenciando aspectos que ocorrem em ambos modelos.

A análise comparativa de semelhanças foi dada a partir da observação dos elementos dos modelos conceituais dos instrumentos, com base no que é considerado um modelo de observação de princípios construído por Campos (2004), a saber: (a) método de raciocínio; (b) objeto de representação; (c) relações entre os objetos e; (d)

formas de representação gráfica. Estes princípios são observados aqui a partir do enfoque da Ciência da Informação e da Ciência da Computação.

O método de raciocínio pretende compreender a sistematização utilizada de *como olhar o domínio*, compreendendo a construção de modelos a partir dos métodos dedutivo e/ou indutivo. O objeto de representação é considerado, em geral, como "a menor unidade de manipulação/representação de um dado contexto" (CAMPOS, 2004, p. 26). As relações entre os objetos permitem que seja observada a estrutura do contexto em que os objetos estão inseridos, sendo possível identificar tipos de relações e como elas ocorrem entre os objetos. As formas representação gráfica permitem que o modelo conceitual seja visto como "um espaço comunicacional em que transpomos o mundo fenomenal para um espaço de representação" (CAMPOS, 2004, p. 31).

3. Aspectos comparáveis entre os modelos conceituais de tesauros conceituais e ontologias de fundamentação

No que diz respeito ao método de raciocínio, a construção de tesauros conceituais conta, basicamente, com o aporte de duas teorias na Ciência da Informação: a Teoria da Classificação Facetada e a Teoria do Conceito. A Teoria da Classificação Facetada visa o estabelecimento de categorias gerais a partir do olhar sobre um domínio, deixando a compreensão dos objetos que as constituem para um momento posterior. A Teoria do Conceito, por outro lado, compreende um modo analítico-sintético de conhecer o domínio, sendo "uma metodologia híbrida [...] agregando [o método dedutivo e indutivo] em um exercício de pensar o particular como um todo e o todo possuindo particulares" (CAMPOS, 2004). Deste modo, entende-se que o domínio apresentará categorias ao final da análise dos conceitos, embora não se conheça inicialmente quais são, chegando-se a elas a partir da análise dos conceitos.

A ontologia de fundamentação se utiliza da indução como método de raciocínio, ou seja, parte da observação dos objetos no mundo (particulares/individuais) para chegar aos universais. Por outro lado, "apesar de possuir princípios para descrição de metaníveis de objetos em um domínio (universais), não utiliza esta classificação como um mecanismo inicial para a organização dos objetos em um contexto" (CAMPOS, 2004, p. 26). Isto permite que a observação dos elementos traga a tona uma estrutura conceitual que revele a real constituição dos mesmos bem como suas relações, já que a partir de uma perspectiva filosófica realista o modelo conceitual gerado é um modelo da realidade.

No que se refere ao objeto de representação, ele é, segundo mostram os estudos realizados, a menor unidade de representação de um contexto. A Ciência da Informação, a partir da Teoria do Conceito, admite a existência conceitos propriamente ditos, sendo este composto pelo referente – o objeto –, suas características e um nome que o designa. Este referente é um objeto no mundo, alguma coisa que realmente existe, sendo classificado como objeto geral ou individual. A Ciência da Computação, a partir da ontologia formal, os objetos, ou particulares, são classificados inicialmente como *endurants* (contínuos) ou *perdurants* (ocorrentes). Os *endurants* são objetos/entidades, enquanto o *perdurants* são eventos/ações. Apesar dos tesauros conceituais não possuírem tal classificação, os conceitos que os constituem são também objetos/entidades, eventos/ações, entre outras categorias de conceitos.

A ontologia de fundamentação está pautada no trabalhar com objetos de representação a partir de uma visão Aristotélica de mundo, estabelecendo a existência de categorias gerais que podem ser usadas de forma a estruturar modelos da realidade, sendo, assim, são passíveis de representação. Deste modo, o modelo formal construído permite o "raciocínio" sobre estes elementos.

No que tange às relações entre objetos, para a construção de modelos conceituais de tesauros conceituais, os conceitos estão relacionados entre si porque existem características comuns entre eles. As características são, assim, essenciais para a construção de relações e o posicionamento do conceito em um sistema de conceitos. Essas características permitem que seja observada a essência do conceito, uma vez que descrição de características essenciais de um objeto permitem sua identificação conceitual, formando, como ressalta Campos (2004), a estrutura conceitual do contexto.

As relações existem tanto em tesauros conceituais quanto em ontologias de fundamentação. Aqui esboçaremos um comparativo de forma a caracterizar as relações existentes em tesauros conceituais que também são previstas em ontologias de fundamentação, sem, no entanto, deixar de perceber que as relações existentes em ontologias de fundamentação são de uma variedade extremamente maior. Deste modo, apresentaremos as relações propostas por Campos (2004) para a modelagem de domínios de conhecimento utilizadas em tesauros conceituais que podem apresentar semelhança com relações na ontologia de fundamentação, a saber: (a) relação categorial; (b) relação hierárquica; (c) relação partitiva e; (d) relação funcional-sintagmática.

A relação categorial é apresentada na construção de tesauros através da relação formal-categorial na Teoria do Conceito. Esta relação toma por base o referente escolhido, impondo-lhe um processo de categorização e permitindo, assim, que seja montada a estrutura do domínio, conferindo estabilidade e flexibilidade a esta estrutura. Nota-se que este processo é substancialmente diferente do utilizado na Teoria da Classificação Facetada, onde as categorias são definidas *a priori* e os elementos "encaixados" nestas categorias. Na ontologia de fundamentação, a observação parte dos objetos e a partir deles são estabelecidas relações. Nesse processo, "a categoria, especificamente, é considerada uma classe de nível mais amplo, tendo como função possibilitar uma classificação geral do domínio em questão" (CAMPOS, 2004, p. 28).

A relação hierárquica é compreendida por conceitos de mesma natureza, ou seja, aqueles elementos que já estão agrupados em determinada categoria. Em tesauros conceituais a relação hierárquica compreende as relações gênero/espécie e a relação lateral (conceitos em renque). Na ontologia de fundamentação, novamente por sua base estar situada na ontologia formal, a relação de gênero/espécie permite organizar taxonomicamente a estrutura do domínio. Neste processo, como lembra Campos (2004), pode ser observada a questão de identidade dos objetos, como forma de verificação da natureza dos mesmos. Pode-se ressaltar também as noções de dependência e rigidez[1].

Outro tipo de relação é a relação partitiva ou parte-todo. Como o próprio nome denota, esta relação compreende a ligação entre o todo e suas partes e a relação das

[1] Dependência: versa sobre a existência de uma entidade espécie estar condicionada a existência de uma entidade gênero. Rigidez: responde pela entidade ser a mesma ao longo do tempo (essência), mesmo sofrendo alterações.

partes entre si. Basicamente este é o entendimento sobre a relação partitiva na construção de tesauros conceituais. A ontologia de fundamentação, por sua vez, compreende o estudo aprofundado desta relação, destinando uma área da Filosofia para estudar especificamente os relacionamentos entre o todo e as partes de uma entidade.

A relação funcional-sintagmática pode ser reconhecida como uma relação que torna "evidente uma determinada demanda, ou função, entre os objetos no mundo fenomenal, não objetivando explicitar o objeto e suas propriedades" (CAMPOS, 2004, p. 30), ou seja, esse tipo de relação é conceitualmente orientada a processos ou operações (DAHLBERG, 1978b). A ontologia de fundamentação tem, na ontologia formal, subsídios para que esta seja trabalhada através da noção de dependência a ligação entre os conceitos, explicitando a "dependência existencial, envolvendo indivíduos específicos pertencentes a classes diferentes" (CAMPOS, 2004, p. 30). Assim, embora não nomeada desta forma, a ontologia de fundamentação possui uma vasta tipologia destas relações.

Chegando, por fim, às formas de representação, pode-se verificar que a Ciência da Informação destina teorias e metodologias consistentes e utilizadas desde muito tempo para a modelagem de domínios, mas as possibilidades de manifestações gráficas não são desenvolvidas. A ontologia de fundamentação tem explorando o ferramental tecnológico para constituição taxonômica de elementos que compõem um domínio, desenvolvendo aparatos capazes de projetar visualmente a constituição do domínio.

Uma vez que nosso intuito é estabelecer os elementos existentes nos modelos conceituais dos instrumentos analisados, não cabe analisar a forma de representação gráfica desenvolvida por uma ou outra área, mas compreender que a Ciência da Computação, a partir da utilização de seu arcabouço de tecnologia da informação, está largos passos à frente da Ciência da Informação nesse processo de desenvolvimento. Este é, sem dúvida, afetado pelo domínio do ferramental de desenvolvimento tecnológico daquela área.

4. Algumas considerações finais

Observa-se que a Ciência da Informação dispõe de bases teóricas e metodológicas próprias para a construção de instrumentos terminológicos, como tesauros conceituais, o que constitui um arcabouço sólido de conhecimentos, capaz de permitir que seja criada uma teoria independente sobre um domínio. Este arcabouço está posto na Teoria do Conceito, a qual permite perceber o domínio a partir de uma análise analítico-sintética, e na Teoria da Classificação Facetada, a qual permite estabelecer categorias gerais de domínio, bem como regras para que isso seja feito de forma válida.

As ontologias de fundamentação, por outro lado, detêm forte fundamentação da Filosofia e das Ciências Cognitivas, permitindo que a estrutura real de um domínio, seu compromisso ontológico, seja representada de forma fiel, clara e consistente. Isso permite que a representação realizada detenha uma semântica baseada no mundo real, restringindo interpretações sobre seus conceitos. Deste modo, as ontologias de fundamentação, permitem que a construção de uma teoria sobre o domínio possibilite testar e validar um modelo conceitual.

Assim, um tesauro, desenvolvido a partir de abordagens da Biblioteconomia e Ciência da Informação, tem sua organização semântica através de relacionamentos e da restrição dos significados dos termos, fazendo com que estes sejam utilizados de forma unívoca. As ontologias, por outro lado, oriundas da engenharia informática e tendo por base a ontologia formal, apresentam relações de maior variedade, também permitindo a representação de determinado domínio. Essa representação é definida formalmente, sendo possível observar a estrutura conceitual (hierarquia) do domínio e receber respostas do sistema a partir de um esquema de inferências. Isto permite que as ontologias possuam maior teor semântico no que diz respeito as suas relações, evitando, quando bem projetadas e filosoficamente bem fundamentadas, diversas inconsistências conceituais.

Por fim, parece não restar dúvidas que um dos grandes fatores de diferenciação entre tesauros (conceituais) e ontologias (de fundamentação) é o entorno digital em que as últimas são desenvolvidas. Embora a utilização de tesauros conceituais seja possível em ambientes digitais, sua utilização é estática, devendo o usuário realizar consultas ao sistema através de assertivas. As ontologias, de outra maneira, são um tipo de sistema capaz de responder questões formuladas pelos usuários. O sistema, então, é capaz de realizar inferências, desde que os elementos conceituais que fazem parte da pergunta estejam em sua base de conhecimento. Ou seja, o fato de as ontologias serem oriundas de um meio computacional, permite que a automatização conferida pelo meio lhe sustente a capacidade de, por exemplo, realizar inferências tendo por base as restrições impostas, percorrendo regras válidas acionadas por meio de complexos axiomas, respondendo questões propostas.

References

CAMPOS, M. L. A. (2004) Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. *Ciência da Informação*, Brasília, v. 33, n. 1, p. 22-32.

COLINO, C. (2002) El metodo comparativo. In: REYES, R. (Dir.). *Diccionario Crítico de Ciencias Sociales.* Madrid: Universidad Complutense.

DAHLBERG, I. (1978a) A referent-oriented analytical concept theory of interconcept. International Classification, *Frankfurt*, v. 5, n. 3, p. 142-150.

DAHLBERG, I. (1978b) *Ontical structures and universal classification*. Bangalore: Sarada Ranganathan Endowment. 64 p.

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

RANGANTHAN, S. R. (1967) *Prolegomena to library classification*. Bombay: Asia Publishing House.