

# Methodological aspects concerning the reuse of ontologies: a study based on genomic annotations in the domain of trypanosomatides

DOI: 10.3395/reciis.v3i1.243en



*Maria Luiza  
de Almeida  
Campos*

Institute of Arts and Social  
Communication, Federal  
Fluminense University, Nite-  
rói, Brazil  
marialuizalmeida@gmail.com



*Maria Luiza  
Machado  
Campos*

Institute of Mathematics,  
Rio de Janeiro Federal Uni-  
versity, Rio de Janeiro, Brazil  
mluiza@pq.cnpq.br

## *Alberto M. R. Dávila*

Oswaldo Cruz Institute, Oswaldo Cruz Foundation,  
Rio de Janeiro, Brazil  
davila@fiocruz.br

## *Linair Maria Campos*

Brazilian Institute of Information in Science and  
Technology, Federal Fluminense University, Niterói,  
Brazil  
linair@hotmail.com

## *Hagar Espanha Gomes*

Federal Fluminense University, Niterói, Brazil  
hagar.espanha@terra.com.br

## *Laura Lira*

Brazilian Institute of Information in Science and  
Technology, Federal Fluminense University, Niterói,  
Brazil  
llira@gbl.com.br

## Abstract

In the last years ontology development has grown considerably, suggesting the maturing of the efforts towards the development of standard vocabularies, specially in Biomedicine, which can be considered as a complex inter and multidisciplinary domain. However, despite the existence of many methodological approaches and best practices to guide the structuring of the ontology hierarchy and its relations, not much is explained about the methods and techniques used to analyze the domain in order to obtain a list of its relevant concepts and to establish its scope, specially if ontology reuse is desired. In this sense, our goal is to present the contributions of Information Science and Computer Science that can be applied to ontology reuse, as a methodological step towards knowledge acquisition. On doing so, we hope to contribute to the enhancement of ontology aligning and mapping mechanisms in the domain of Trypanosomatids

## Keywords

ontology reuse; ontology alignment; language compatibility; Trypanosomatids; knowledge acquisition

## Introduction

Initiatives of the international scientific community in the field of genomics for the last years have led to an explosive growth of biological information, which is being continuously generated every day (HGP 2003). The initial concern was, therefore, the creation and maintenance of a database to store biological data. As genomic databases are filled and genomes sequenced, studies gradually shift their focus from genome mapping to the analysis of the broad range of information resulting from the functional characterization of genes by means of Molecular Biology and Bioinformatics. It becomes essential to create an interface among the data obtained through various research projects around the world on the interrelation of enzymes, genes, chemical components, diseases, species, cell types, organs, etc. (Mendes 2005). In order for these teams and/or institutions to exchange scientific resources, it is necessary to find a common method to describe and access these resources, so as to facilitate their search, integration, and reuse.

Hence it is important to consider the relevance of management, description and organization of scientific resources in a digital medium for the research in the field of Bioinformatics. It should be highlighted that Bioinformatics is an interdisciplinary field comprised of Biology, Computer Sciences and Information Technology, whose purpose is to enable the discovery of new biological introspections, as well as to create a global perspective where the unified principles of biology can be distinguished (Belloze 2007)

The large amount of data being accumulated in the various databases around the world needs to be annotated and interpreted with the utilization of available genomic data. For this purpose, the various projects interested in exchanging and integrating information must adopt standards for the annotation of data, in order to consistently enable information retrieval. Ontologies play an essential role in this integration, enabling the semantic interoperability of heterogeneous distributed systems. Such is the case of initiatives involving international consortiums (Campos 2007).

The Open Biological Ontologies (OBO) Ontology Library (OBO 2005) is a terminology repository developed for shared utilization among several biological and medical domains. Despite the fact that it is called an *ontology* repository, vocabularies can actually be defined in several ways, such as: controlled vocabularies, glossaries, and ontologies per se. Additionally, some vocabularies can be generic to the point of being applicable to all organisms, while others have terms which are restricted to taxonomic groups such as flies, fungi, yeast, or fish. Among OBO's most disseminated vocabularies, we can highlight Gene Ontology (GO) (Gene Ontology Consortium 2001). GO is comprised of terms concerning three main categories: cellular components, biological processes, and molecular functions, non-dependent on organism species (Ashburner 2002).

Brazil – or, more specifically, the activities of the field of genomic scientific applications – is developing

a project called “Genome and Comparative Transcriptome: a Bioinformatics consortium for the development of a Web platform and integrated databases” under the supervision of Fiocruz. One of the main targets of this project is to provide an environment that is able to offer semantic information on scientific resources, such as data and software in the Bioinformatics area, and to enable the shared utilization of these resources by the scientific community. GO is being used for annotations on its database.

The implementation of this environment required the formation of a consortium involving Fiocruz and the Federal Universities of Rio de Janeiro and Santa Catarina, with the purpose of creating a Bioinformatics portal and an integrated web platform for analyzing genomes and transcriptomes. The development of capacities and infrastructure in the Bioinformatics area is strategic in Brazil. As a consequence, it is also greatly important for collaboration between the various initiatives of the genome projects, both in Brazil and abroad. Therefore, with the purpose of assisting, optimizing, and disseminating research, a platform called BiowebDG is being progressively implemented, as the result of an equally named consortium, publicly available at: <http://www.biowebdb.org>.

The BioWebDB Consortium, funded by CNPq, is comprised of a group of researchers from the fields of Biology, Bioinformatics, Computing and Information Science, who carry out studies in comparative genomics and genomic databases. Comparative Genomics encompass the analysis and comparison of genomes from different species, with the purpose of achieving a better understanding on how species evolved, or of determining the roles of genes and non-coding regions of the genome by means of such comparisons. Much of the existing information on human genes could be discovered due to the analysis of their correlates in simpler model-organisms, such as mice (HGP 2003).

Studies carried out by the group focus on three main subjects: developing Bioinformatics tools to analyze genomes; analyzing trypanosomatid genomes; developing ontologies; and creating compatibility amongst languages. The consortium's initiative intends to build flexible, intelligent, integrated and friendly platforms, shareable among different data sets and genome projects. In this context, ontologies assume fundamental importance in ensuring semantic organization and information recovery.

The study we are carrying out already points towards some results that allow us to state that no ontologies can be currently identified, either domestically or internationally, which have been developed in accordance with the specific conceptual cut-down – trypanosomatids – to meet the requirements of the groups under Fiocruz's coordination. Despite international initiatives, Gene Ontology does not have concept classes which can fully meet the needs of the studies developed in Brazil. In some instances, it is necessary to investigate the harmonization between terms and their conceptual

content. To this extent, and still as a proposal of the OBO consortium, there exists an encouragement towards the elaboration of specific GO cut-downs (called GO Slims<sup>1</sup>), whose purpose is to provide GO subsets, quite frequently with less detailed hierarchies and focusing on specific organisms.

However, despite the dissemination of languages and tools for the representation and construction of ontologies, their underlying methodologies are of little use, as they usually do not encompass adequate guidelines, either for the identification of concepts and their relations, or for the creation of systematic definitions associated with such concepts. Consequently, tools are of little assistance in guiding users in the ontology construction process, as well as in providing management strategies for the construction of high quality ontologies (Gangemi et al. 1996, Fernández et al. 1997).

This article intends to discuss the issues inherent to the reuse of ontologies, as a methodological step towards the acquisition of knowledge in ontologies, and thus offer tools for mapping and matching ontology terms within the domain of trypanosomatids.

This study is, therefore, organized as follows: section 2 deals with basic aspects of reusing ontologies; section 3 deals with related studies; section 4 provides details preliminary to our proposal for methodological aspects employed in the reuse of ontologies; section 5 presents a discussion about our work and the difficulties we were confronted with. Lastly, section 6 contains final considerations.

## Ontology reuse

As several studies have presented, ontology (Gruber 1993, Guarino 1993, 1998, Vickery 1997, Swartout & Tate 1999, Corazzon 2000, Smith 2002) as a knowledge representation tool appeared in the 1990s, within the Artificial Intelligence field. For Artificial Intelligence systems, what exists is what can be represented. When knowledge of a domain is represented in a declarative language, the group of objects that can be represented is called the domain of discourse. Ontologies appeared with the purpose of describing data handled by applications, by means of defining a group of terms that could represent domains and tasks that those applications should execute.

An ontology is, hence, a group of standardized concepts, terms, and definitions accepted by a specific community. Gruber's (Gruber 1993) is the most common definition of ontology: "an ontology is the specification of a conceptualization".

A conceptualization is an abstraction, a simplified view of a world that is represented to meet one or more of the following purposes: "allowing multiple agents to share their knowledge; helping individuals to better understand a certain field of knowledge; helping individuals achieve a common understanding on a field of knowledge" (Smith & Falbo 1998). In Logics, a conceptualization identifies the object and relations existing in the logical universe (Weinstein 1998).

Ontologies can be reused in many ways, which sometimes results in the creation of an independent ontology based on the concepts of other ontologies (which may be extended and adapted), and sometimes preserves the original ontologies. The second case is the approach we employ, which is called *ontology matching*.

Matching produces different results from merging and integrating: instead of managing an additional ontology, which is the result of combining reused ontologies, it keeps reused ontologies unchanged in their original locations, but generates a set of links among them. These links contain a variety of information on how to make reused ontologies compatible, and are expressed on a separate persistent (physically existing) model.

A set of links expressed in a persistent model produced by means of the matching process is a *mapping* between ontologies. Information contained in the mapping will depend on the type of semantic relationship existing among elements and on the type of formalism used in the ontology to represent its semantics. For example, two elements may be similar (to varying degrees), or one can be a part of the other, or they may have some other kind of relation that is identified with the help of a domain specialist.

Similarity mappings can express varying degrees of similarity (Felicísimo & Breitman 2004, Kalfoglou & Schorlemmer 2003, Aleksovski et al. 2006, Su 2004). Usually several factors are taken into account in order to determine the degree of similarity, such as: linguistic similarity between terms; compatibility of their attributes; term's position within the hierarchic structure, among others. One of the issues of mapping concerns how to find candidates. For more details on these issues, De Bruijn et al. (2006) have carried out a consistent survey on types of conflicts that appear when mapping ontologies.

Another aspect involving matching concerns the type of technique employed to estimate candidates. It can be based, among other aspects: (i) on similarities between term names; (ii) on the ontology structure, such as, for instance, it may consider the terms' position within the hierarchical structure of ontologies under comparison, or their partitive relations, or other types of relations that are similarly used in compared ontologies (Euzenat & Shvaiko 2007); (iii) on the addition of supplementary knowledge, such as, for instance, on information from another ontology or vocabulary with a concept hierarchy, such as Wordnet (Miller 1990), which may be used, for instance, to search for synonyms, or to compare the distance between the positions of terms of the ontologies being mapped against this other ontology (Reynaud & Safar 2007, Sabou et al. 2006).

## Studies related to the reuse of ontologies

Literature on the reuse of ontologies minutely explores the various aspects involved from an operational point of view, that is, what needs to be done or arranged, and problems that arise within this context. Regarding methodological aspects on how to reuse them, what we

more often find concerns computational aspects, such as, for instance, which algorithms are most effective to promote compatibility among ontologies, both regarding the accuracy and the speed of their results (Noy & Musen 2000).

Some authors even propose more general tasks that are necessary in the reuse process. Gangemi, Steve and Giancomelli (Ganemi et al. 1996), for instance, state that it is necessary to identify the basic terms and their necessary and sufficient definitions in textual format. However, they provide no suggestion on how to perform such identification, or on which principles should be adopted to build the definitions. The more comprehensive view of Pinto and Martins (2001), on the other hand, suggests that the reuse process starts during the selection of ontologies to be reused. No further details are given, however, on how to perform such tasks.

Ours studies have been pointing towards the importance of investigation within the sphere of studies on Building Language Compatibility in the Information Science domain. We consider that they will provide us with theoretical and methodological guidelines for the reuse of ontologies (Campos 2005).

### Semantic aspects of reuse concerning the compatibilization of vocabularies

One of the aspects of reuse is the compatibility among reused vocabularies. It should be highlighted that the word compatibility has a very specific definition in the field of Computing Science. It concerns the ability of computers of various types to run software developed in a different computer language, without the need to convert it. In this sense, it is important to clarify that our utilization of the term is defined within the boundaries of Information Science, and is a seminal study of that field, with theorists such as Soergel (1982), Dahlberg (1981), Neville (1970,1972), and Glushkov (1978), Campos (2006).

For Glushkov and others (1978), compatibility is the level of similarity between two languages, where there exists a concept of degrees of compatibility and a distinction between semantic and linguistic compatibility.

Two methods distinctly stand out among the others used for converting and creating compatibility between languages based on the integration of vocabularies. These are Neville's thesaurus reconciliation method (1970, 1972), and Dahlberg's concept correlation matrix (1981, 1983).

Neville's method is based on the principle that concepts (the conceptual contents of descriptors, which are expressed by the definitions), and not descriptors alone, must be made compatible. This method suggests an intermediate language approach, based on the numeric coding of concepts, which enables the establishment of a conceptual equivalence of descriptors of different languages.

The method suggested by Dahlberg (1983) is based on the construction of a concept compatibility matrix by

means of his analytic and synthetic method. The concept compatibility matrix is a mapping of the semantic potentiality of the languages being examined. It provides the results of the language compatibility analysis from the semantic and structural points of view. According to Dahlberg, compatibility between languages is comprised of three phases, as follows: 1. concept coincidence – when two concepts combine their characteristics – degree of equivalence; 2. Concept correspondence – two concepts combine most of their characteristics – similarity; 3. concept correlation – two concepts correlate by means of mathematic symbols, thus establishing a correlation degree.

Compatibilization, however, implies that vocabularies must possess some degree of compatibility and that the more compatible they are, the more accurate and easy is their compatibilization. In order to be more compatible, vocabularies should ideally follow rules that provide guidelines for a more even and standardized construction. Lancaster (1986) had already noticed this issue concerning the construction of thesauri:

“By creating a structural compatibility among vocabularies, norms facilitate the conversion of one vocabulary into another. Therefore, two thesauri following ISO norms for the construction of thesauri are likely more easily reconciled than two thesauri built on different principles. Moreover, such norms promote compatibility in a general sense: Once an information service user is familiar with a thesaurus, it would be easier for them to convert information into another thesaurus, built under the same rules.” (Lancaster 1986, p 212).

It is important to highlight that most of the times matching proposals explore the compatibilization of similar-meaning terms, assuming that it is necessary to maintain different vocabularies even if they refer to subjects that possess a certain degree of overlapping. This way of conceiving ontologies as vocabularies that express different views of the same domain is not consensual, though. Especially regarding the Biomedical field, where vocabularies have complex themes.

Some authors, such as N. Guarino and Barry Smith suggest slightly different proposals, although both focus on the standardization of ontologies based on an examination of the classification of their concepts and relations.

Some of the many studies carried out by Guarino (1998a) explore the semantic and formal nature of concepts of an ontology. In practice, Guarino's *Formal Ontology* can be defined as the theory of *a priori* distinctions concerning: worldly entities of the world (physical objects, events, regions, amounts of matter); meta-level categories to model the world (concepts, properties, qualities, states, roles and parts). Guarino, however, accepts the creation of several not necessarily complementary views of a same domain, which he calls “possible worlds”.

Barry Smith (Smith et al. 2007), on the other hand, is inspired by the Aristotelian Theory of Classes to suggest a jointly-developed set of axioms and definitions to

be applied in the Biomedical domain. Although Smith's view of ontology categorization is philosophically close to Guarino's (Bateman & Farrar 2004), Smith, as opposed to Guarino, advocates the idea that there is only one "possible world", albeit with different, orthogonal, complementary views. To Smith, ontologies:

"(i) must be developed by means of a joint-effort, (ii) employ common relations, which are defined in a non-ambiguous manner, (iii)... (iv) have a clearly defined subject (in such a way that an ontology concerning cellular components, for instance, does not include terms such as 'database' or 'whole')..."(SMITH et al., 2007, p.2).

In addition to investigating the approaches of theorists such as Smith and Guarino, our research is finding support in studies carried out in the field of language compatibilization, in the sphere of Information Science. Especially in those theories more specifically linked to the representation of concept systems, where there is a solid theoretical foundation for the elaboration of European-based languages, which will provide a semantic base for integration, such as: S. R. Ranganathan's Faceted Classification Theory (Ranganathan 1967), and I. Dahlberg's Concept Theory (Dahlberg 1978a,b, 1983), which allow the representation of knowledge domains. Due to the focus of this article, we will not go into greater details concerning these theories. We will, however, briefly discuss Ranganathan's contribution, once it is employed in the current stage of our work, as shown in section 4.

Ranganathan elaborates a series of principles, which intend to allow the concepts of a knowledge domain to be systemically structured. That is, concepts are organized in ranks and chains, which are in turn structured in comprehensive classes, which are the facets, and the latter are organized within a given fundamental category. The grouping of all categories comprise a concept system for a given subject area, and each concept within the category is also the manifestation of that category (Campos 2001). Categorization is a process that requires the domain to be viewed in a deductive way, that is, it requires the determination of classes with greater comprehensiveness within the chosen subject. The categorization exercise can clearly show the thematic domain of the ontology, and consequently set the grounds for selecting terms from the sources from where they will be taken.

It is in this environment that the foundations where its theory is based can assist in cutting down the domain for the elaboration of ontologies, and ultimately for the construction of concept models. His [Meta] Category postulate, of special interest to our study, suggests the existence of five basic categories, which can be used to cut down subject universes into comprehensive classes. Regardless of which categories are employed to consider the structuring of a domain (five, less, or more), it is important to consider Ranganathan's idea that they encompass concepts when compatibilizing vocabularies, once they enable an expansion of the semantics of the nature of classes. This outlook is being explored during

the initial phase of our experimentation. We expect to be able to explore other IS contributions previously mentioned in this section.

As we can see, the organization of knowledge domains is receiving attention, both in Information Science and in Computer Science, in a very independent way, sometimes specific in certain aspects, such as the hierarchical organization of concepts, or the efficiency of computer algorithms. Our proposal intends to bridge the gap between these areas, and to broaden and integrate, whenever possible and appropriate, discussions on proposals for domain organization within the sphere of ontology reuse.

In this panorama, the re-examination of literature seems to point toward a lack of comprehensive and detailed proposals concerning issues that precede and are the foundations of ontology reuse itself, placing them within a context that enables the comprehension of their origin, motivation, purpose and application scenarios. Criteria for ontology selection are not limited to identification of ontology characteristics that will be analyzed. They must consider not only principles that must guide the analysis, but also principles that will outline the context where ontologies will be reused, both from the point of view of their immediate applicability, and of the environment where they are inserted. We assume that a more consistent and accurate reuse can be achieved by means of the identification and specification of these principles.

## **Ontology compatibilization: applicability in the domain of trypanosomatides**

The consummation of our proposal occurs, as previously mentioned, within the sphere of BioWebDB Consortium's projects, joining both theoretical and experimental efforts, employing an interdisciplinary team, and being supported by researchers from various institutions<sup>2</sup>. In this scenario, this section will deal with the first experiments concerning ontology compatibility under the reuse concept. Our experiment has so far approached two main issues: the methodology employed in composing the term sample, and the reuse approach adopted for application in the selected sample.

## **Vocabulary selection within the domain of trypanosomatides**

Methodological studies carried out within the domain of Information Science to support the survey of terms that compose the units of a given knowledge domain have been examined by many researchers (Sorgel 1982, Lancaster 1986, Dahlberg 1978b, Hjørland 2002). These studies provide systematic guidelines that have been examined – in the context of this study – for a preliminary analysis of the domain. Support provided by these theoretical contributions and by others from the Social Sciences (Latour 1997) have allowed us to elaborate an initial draft of thematic groupings of the

domain of Trypanosomatides in the Molecular Biology Laboratory of Trypanosomatides and Phlebotomines of the Oswaldo Cruz Institute (IOC).

Latour's player-network theory (1997) establishes that science must be studied within the practices of scientists, including the man-machine and society relation. Science occurs on laboratory benches, defining their content and the whole context where these players operate in the social environment along the process of acting. In this sense, it is essential that we have an outlook on the domain of interest out of our active participation within it. Therefore we have been participating in a series of seminars and interviews, which help better understand this domain.

Hjørland (2002) states that Information Science has informational resources that must be identified, described, organized and advertised in order to meet specific purposes, and that it can benefit from considering the analytic view of the domain by means of different approaches, such as: the analysis of specialized texts, assessment of computing tools, user analysis, among others.

Bearing these perspectives in mind, the analysis of the researcher's domain initially followed a mapping criterion, both concerning activities carried out in the laboratory and literature, with the intention of identifying, on the one hand, a set of ontologies where reuse tools could be employed and, on the other hand, a set of terms to serve as a sample basis for compatibilization activities, as we will see below.

An assessment of subjects and ontologies of interest (besides GO) was carried out based on the literature resulting from the studies undertaken within the scope of the Laboratory. Initially ten large thematic groups were identified: Protists, functional and Systems Biology, Molecular Biology and Genomics, evolutionary molecular Genetics, comparative Genomics, Phylogeny, Bioinformatics, Diseases, Metagenomics, Drug targets; each of them with subsets that we are currently specifying and validating<sup>3</sup>.

Ontologies related with thematic groupings within the scope of OBO were thus mapped. A group of seven ontologies was classified as being of interest: NCBI organismal classification, Pathway, Sequence types and features (SO), Brenda tissue / enzyme source, Event (INOH pathwayontology), Multiple matching and System biology (OBO 2005). These will be used as domains so that we can identify classes within the sphere of the domain of Trypanosomatides.

On the other hand, a set of 800 terms resulting from the genomic annotations existing in the GARS system (Davila et al. 2005)<sup>4</sup>, annotated according to the Gene Ontology (GO) and resulting from research (Wagner 2006), within the sphere of the functional genomics of Trypanosomatides, especially concerning the *Trypanosoma rangeli* species were used for comparison with selected OBO ontologies, so as to provide us with several hierarchies of parent and child terms for each term found, with their corresponding definitions and partitive relations,

when applicable. A software application was developed for this purpose, not only to extract, but also to convert the language of the ontologies employed (originally in OBO format) into OWL language (Web Ontology Language) (OWL 2008), so as to facilitate future inferences and computer handling, since this material along with the hierarchy of ontologies will be used as a sample for reuse experiments.

### Approach used in the reuse

Choosing an approach for the reuse depends, among other factors, on the purpose one wishes to achieve, and on the context where reuse is inserted. Regarding our experimental scenario, our purpose is to describe genomic sequences of trypanosomatides within an integrated view of the genome, transcriptome, proteome and metabolome of such organisms. The following aspects of their usage context must, therefore, be considered: (i) GO, a broad-usage vocabulary in Biomedicine, must not only be reduced in scope for trypanosomatides, but also supplemented with others concerning aspects not covered by it, such as metabolic ways and diseases; (ii) description of these sequences must somehow point toward standardized vocabularies of the area, especially GO, due to the hegemonic employment of these vocabularies in the annotation of genomic resources; (iii) the research group is unable to bear the charges involved in updating the ontologies created, once it relies on scarce resources.

It must be highlighted that, despite the existence of efforts towards the reformulation of OBO ontologies aiming at their factorization into orthogonal, well-defined and organized ontologies, this is not yet the current reality. Therefore, until this initiative becomes a reality it is important to deal with the overlapping of similar subjects and concepts existing across ontologies with different subjects and distinct definitions.

By taking into account the above mentioned factors, we have concluded that matching is the most suitable process for our study. The methodological strategy adopted for matching the ontologies selected in item 4.1 is based on the criterion of semantic compatibilization supported by additional knowledge, the latter being initially obtained from the examination and identification of the nature of first-level ontology classes. This study is carried out under the outlook of fundamental categories, and finds support in the Classification Theory (Ranganathan 1967).

Regarding actual ontology matching execution, we must mention the importance of the support offered by software tools such as Prompt (Noy & Musen 2000), Chimera (Mcguinness et al. 2000) or Fca-Merge (Stumme & Madche 2001), due to the complexity of the task and the possibility of automating some activities. Especially regarding the task of finding candidate terms (matching) for mapping.

In this regard, we intend to investigate whether the application of proposed methodological principles contribute to an increase in the *accuracy* of software tools in

obtaining terms of interest to our experimentation area. With this purpose, we put specific efforts towards the adaptation of a software tool, developed as the result of a graduation project of the Computer Science course of the Federal University of Rio de Janeiro (UFRJ), whose purpose is to match ontologies by means of an algorithm that explores its hierarchical structure and the properties of its classes concerning similarities found in their names (Silva 2008). In addition to ontology structure and name similarities, our adaptation suggests considering the semantic nature of these classes and properties.

### Preliminary results

The current stage of our tests is being executed semi-automatically, within a limited set of 28 terms taken from a randomly selected sample.

A cut-down is made for each OBO ontology, generating ontologies with the ascending and descending hierarchies of the selected terms. Each of these ontologies is then mapped with a GO subset containing hierarchies of the 28 selected terms. Mapping is done with the assistance of the Prompt tool. Tool support is essential to Biomedicine, due to the large amount of terms of its ontologies (some containing over nineteen thousand).

Each mapping suggestion provided by the tool is then manually analyzed, and three aspects are assessed: similarities in term designations, semantic similarity indicating concepts of similar nature (logically related), relations indicating concepts that are not similar, but that may be associated by means of category (logic) relations which are relevant to the domain. In the last case, we are currently attempting to assess the complexity and feasibility of manually executing this task.

The purpose of this experiment is to identify an ideal set of accurately suggested mappings, with the highest amount of usable suggestions. The adopted methodology intends to increase the semantics of handled

ontologies. It is worth noticing that at the current stage of our research, the technique employed to assess the correspondence of terms candidate for mapping (see Figure 1) finds support in the similarity of term designations, in the analysis of ontology structure and of the use of additional knowledge, to be incorporated by means of a high-level formal ontology, specifically elaborated for the given domain.

Based on a preliminary analysis, our results already suggest an increased accuracy when handling false positives, which brings us closer to the ideal set of intended mappings. These are still initial results, and their scope must still be broadened and revised. We can, however, consider that they point towards promising evidences that validate our assumption.

As an example we can mention the mapping of the *excretion* concept, found in GO and Brenda ontologies. In the former, the term refers to a process and means “elimination of excreta by an organism, resulting from metabolic activity”. In the latter, it refers to the product of an activity and means “the matter, such as urine or sweat, excreted by blood, tissues, or organs”. When we map both ontologies through the Prompt tool, it indicates that terms are similar but actually require a semantic analysis.

Similarly, the terms *transporter*, from the MoleculeRole (a branch of INOH) ontology, and *transport*, from GO, also generate false positives in the mapping suggested by Prompt. *Transport*, as in GO, is a process defined as “processes specifically pertinent to the activities of integrated living units: cells, tissues, organs and organisms”. *Transport*, as in MoleculeRole, on the other hand, is a protein defined as “linking specific solutes to be transported that undergoes a series of conformation changes to transfer the linked solute (...)”. Figure 1 shows examples of this type of result, obtained from our preliminary analyses.

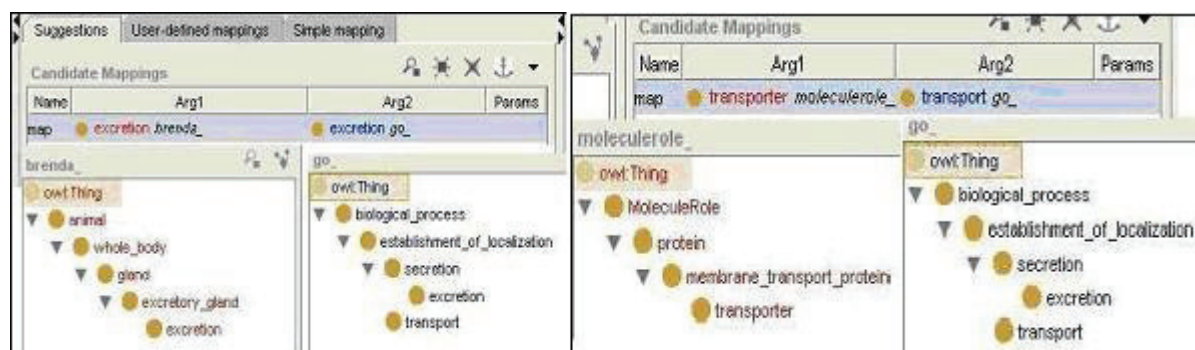


Figure 1 - False positives suggested for mapping by the Prompt tool.

As we can see, these term pairs, despite their linguistic similarity, denote concepts with distinct natures (different fundamental categories). Therefore, they should not have been suggested as mapping candidates due to

conceptual similarity (logic-type relation), as suggested by the Prompt tool.

On the other hand, when linguistic similarity is confronted with a set of predefined category relations, which

can be suggested by the machine for man's validation, allows us to perceive that terms can be associated by means of an ontic-type relation. In the case of Figure 2, we were able to identify the relation between the terms *excretion* (Brenda) and *excretion* (GO) by means of a *process-product* category relation, that is, *excretion* (a matter, in Brenda) is the *product of excretion* (an activity, in GO). Similarly, we identified that *transporter* (a protein, in MoleculeRole) *participates in transport* (a process, in GO).

## Discussion

The Biomedical domain is complex and challenging, even if limited to the study of specific species within a research laboratory.

On the one hand, one has to deal with the human dimension, reflected in the difficulties created by different languages, by complementary knowledges, such as knowledge possessed by a professional working in the field of informatics, by a biologist, and by an information scientist, each possessing their own research bias within the domain and their own level of maturity: from recently-graduated to senior researchers with broad experience in

the field. Each of them has a different view of the domain, and these views must be reconciled within a broader Biomedical perspective, making use of existing efforts.

On the other hand, one has to deal with the technological dimension, which is essential in an area characterized by complexity and by the adoption of vocabularies containing thousands of terms and several problems which are, in fact, standard.

In this context, our experiments point toward a huge room for improvement in the analyzed ontologies, which lack mechanisms able to more accurately integrate them, by considering not only the technological aspects, whether they can be processed by the machine, but also the way man understands them.

We could detect several compatibility problems in all 28 terms we analyzed, as follows: (i) similar concept definitions in different abstraction levels; (ii) terms with similar denominations and distinct meanings; (iii) terms showing non-explicit interrelations, among others. These problems are being employed in our research as subsidies for the improvement of semantic accuracy of ontologies, as shown in Table 1.

**Table 1 – Subsidies for the improvement of semantic accuracy of analyzed ontologies**

| Nature | Input obtained after analysing mapped terms  |
|--------|--|
| (i)    | The term <i>transporter</i> on <i>system biologys</i> ontology is generically defined as: "Participating entity that facilitates the movement of another physical entity from a defined subset of the physical environment (...) to another". In MoleculeRole ontology, the definition of participating entity and physical entity is specified for protein and solute, respectively. On confronting these two definitions, we can notice that the use of definition patterns may bring more precision to the formulation and comprehension of concepts. For example, the first definition mentioned above could be used as a definition pattern to be followed by others, more specific, like the second one. |
| (ii)   | The study of terms definitions has confirmed, up to the present moment, the following types of fundamental categories: biological process, molecular function, event, biological component, chemical component, phenotype.   |
| (iii)  | Term definitions indicate, up to the present moment, the following types of categorial relation (which have not been found on relation ontology): process-biological component, process-firing event, process-input, process-product - on the last two, the input and the product are both chemical components (organical ou inorganical).   |

Some difficulties were found in obtaining preliminary results with the ontologies mentioned in this article, such as: the size of some ontologies generated errors in the applications used to handle them, which required them to undergo a previous treatment; the lack of mechanisms that enable a search for similar terms to be simultaneously performed in several ontologies, which required the elaboration of specific software tools; the lack of standardization and documentation concerning the methodology used in the construction of area specific ontologies, which called for a methodic and often unsuccessful search for explanatory materials in various and decentralized sources; the complexity of the domain, which required a large amount

of reading, courses and seminars in the Biomedical field, in order to better understand the context where ontologies are and their definitions are inserted, and to facilitate conversations with field specialists.

## Final considerations

Biomedical research is characterized by a large amount of data, by the complexity of the subject, and by a growing number of vocabularies that attempt to describe and organize related scientific resources.

Such vocabularies are mostly built to meet interests that do not always meet the needs of the research carried out in Brazil. Additionally, they contain structural prob-



lems that suggest an absence of methodologies focusing on their development.

However, due to the high degree of complexity of the domain, the high costs involved in the construction of such vocabularies, and their wide adoption by the Biomedical community, their reuse has to be considered in the elaboration of vocabularies more suitable for domestic research.

Given the context, our intention is to discuss the issues inherent to the reutilization of ontologies, particularly those related to the mapping and matching of ontology terms within the domain of trypanosomatids.

As a starting point, we are carrying out experiments that focus on acquiring the domain knowledge, intending to provide support for the theoretical bases we are examining. As a preliminary result, we wish to highlight a set of 28 hierarchies and terms, with their corresponding definitions and partitive relations, that are relevant for the research carried out in the Molecular Biology Laboratory of Trypanosomatids and Phlebotomines of Fiocruz's IOC. These samples, whose subjects are complementary and overlapping to some extent, is an important tool for essays involving issues concerning the reuse of ontologies, and is being explored for compatibility assessment and concept definition purposes.

A preliminary examination of such hierarchies produces results that already validate our proposal for the semantic enrichment of ontologies, based on the identification of fundamental categories as an important factor for the increase in the accuracy of software tools, primarily utilized in the Biomedical field.

Future studies, which have already been outlined, intend to deepen the domain analysis by means of semi-automatically handling previously selected area-specific literature; of employing other contributions from Information Science in vocabulary compatibilization; and of using high-level ontologies to ponder the relations among ontologies with complementary themes.

## Notes

1. GO Slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They are generally used to describe a specific organism or specific biological aspects only (e.g., cellular locations only). Several GO Slims are currently available, and they can be obtained from the Gene Ontology consortium website.

2. These studies are the preliminary results of two research projects, supported by CNPq, as follows: "Ontology Integration: the bioinformatics domain and issues involving terminological compatibilization", in the Information Science field; "Comparative genome and transcriptome", in the Computing Science field. In addition to the projects, these are subjects addressed by the researches of two students from the doctorate course Post-Graduate Program in Information Science UFF / IBICT. In all researches, the empirical field of action is linked to genomic studies within the sphere of the BioWebDB consortium.

3. We are currently testing some automatic extraction tools in order to assess terms by means of a non-manual methodology.

4. System developed at Fiocruz for analysis and annotation of genomic resources.

## Bibliographic references

Aleksovski Z, ten Kate W, van Harmelen F. Exploiting the Structure of Background Knowledge Used in Ontology Matching. In: Workshop on Ontology Matching at ISWC, 2006.

Ashburner M, Lewis S. On Ontologies for Biologists: the gene ontology – uncoupling the web. In: *Silico Biology*, Novartis Found Symposium, 2002, p. 66-83.

Bateman J, Farrar S. Towards a Generic Foundation for Spatial Ontology. In: *Formal Ontology In Information Systems: Proceedings of the Third International Conference (FOIS-2004)*, 2004, p. 237-248.

Belloze KT. Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular. 2007. 147 f. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2007.

Campos ML. A Linguagem documentária: teorias que fundamentam sua elaboração. Niterói, RJ: Eduff, 2001.

Campos MLA. Integração de Ontologias: o domínio da bioinformática. RECIIS. 2007; 1:117-121.

Campos MLA. Integração de ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica. (Projeto de Pesquisa submetido ao CNPq no período de 2005 a 2008). Universidade Federal Fluminense- Departamento de Ciência da Informação, 2005a.

Campos MLA. Integração de ontologias: o domínio da bioinformática e a problemática da compatibilização terminológica. In: VII Enancib, 2006, Marília. Anais... Marília, 2006.

Campos MLM, Campos MLA, Campos LM. Web semântica e a gestão de conteúdos informacionais. In: Carlos H. Marcondes; Hélio Kuramoto; Lídia Brandão Toutain; Luís Sayão. (Org.). *Bibliotecas digitais: saberes e práticas*. Salvador, BA; Brasília: EDUFBA; IBICT, 2005, p. 55-75.

Corazzon R. Ontology: a resource guide for philosophers. 2000. Disponível em: <<http://www.formalontology.it>>. Acesso em: 1 jul. 2006.

Dahlberg I. A Referent-oriented analytical concept theory of interconcept. *International Classification*, Frankfurt, 1978a; 5(3):142-150.

Dahlberg I. *Ontical structures and universal classification*. Bangalore: Sarada Ranganthan Endowment, 1978b. 64 p.

- Dahlberg I. Towards establishment of compatibility between indexing languages. *Internacional Classification*. 1981; 8(2): 88-91.
- Dahlberg I. Conceptual compatibility of ordering systems. *Internacional Classification*. 1983; 10(2):5-8.
- Dávila AMR, Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, Mazzoni CJ, Wagner G, Pires PF, Grisard E C. GARS: Genomic Analysis Resources for Sequence Annotation. *Bioinformatics*. 2005.
- De Bruijn J, Ehrig M, Feier C. Ontology mediation, merging and aligning. In: John Davies, Paul Warren, and Rudi Studer: *Semantic Web Technologies*, John Wiley & Sons, 2006.
- Dervin B. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In: Glazier J, Powell R (editors), *Qualitative research in information management*. Englewood, CO: Libraries Unlimited, 1992. p.61-84.
- Doan A, Madhavan J, Domingos P, Halevy A. Learning to map between ontologies on the semantic web. *Proceedings of the 11th international conference on World Wide Web*, Honolulu, Hawaii, USA, may, 2002, p. 662-673.
- Euzenat J, Shvaiko P. *Ontology matching*. Springer Verlag, Berlin, Heidelberg (Germany), 2007.
- Falbo RA. Integração de conhecimento em um ambiente de desenvolvimento de software. Rio de Janeiro: COPPE/UFRJ, 1998. (Tese apresentada à COPPE/UFRJ para obtenção do grau de Doutor em Ciências (D.Sc.) 81f. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1998.
- Felicíssimo CH, Breitman KK. Taxonomic Ontology Alignment - an Implementation. *Proceedings of the 7th International Workshop on Requirements Engineering*, Tandil, 2004. p. 52-163.
- Fernández M, Gómez-Pérez A, Juristo N. *Methontology: from ontological art towards ontological engineering*. Spring Symposium Series. Stanford. 1997. p. 33-40.
- Gangemi A, Steve G, Giancomelli F. ONIONS: an ontological methodology for taxonomic knowledge integration. *ECAI-96 Workshop on Ontological Engineering*, Budapest, Aug. 13, 1996.
- Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*. 2001; 11(8):1425-1433.
- Glushkov VM, Skorokhod'ko EF, Strongnii AA. Evaluation of the degree of compatibility of information retrieval languages of document retrieval systems. *Autom Doc & Math Ling*. 1978;12(1):18-26.
- GO (org.). Portal da Gene Ontology. Disponível em:** <<http://www.geneontology.org>>, acesso em: 24 abr. 2008.
- Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition*. 1993; 5: 199-220.
- Guarino N. Formal ontology and information systems. In: FOIS '98, 1, 1998, Trento, Italy. *Proceedings Amsterdam: IOS Press; Tokyo: Omsha*, 1998a. p. 3-15.
- Guarino N, Carrara M, Giaretta P. An ontology of meta-level categories. *LADSEB-CNR Int. Rep. 6/93*, Preliminary version, nov. 1993.
- HGP. Human Genome Program, U.S. Department of Energy, Genomics and its Impact on Science and Society: A 2003 Primer, 2003.
- Hjørland B. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of Documentation*. 2002; 58(4): 422– 62.
- Kalfoglou Y, Schorlemmer M. Ontology mapping: the state of the art. *The Knowledge Engineering Review*. 2003; 18(1): 1–31.
- Lancaster FW. *Vocabulary control for information retrieval*. 2nd ed. Arlington, VA: Information Resources Press, 1986.
- Latour B. *Ciência em ação: como seguir cientistas e engenheiros sociedade afora*. São Paulo: Editora Unesp, 1997.
- Mcguinness D, Fikes R, Rice J, Wilder S. The Chimaera Ontology Environment. *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI 2000)*, Austin, Texas, Jul. 30-Aug. 3, 2000.
- Mendes PN. *Uma Abordagem para Construção e Uso no Suporte à Integração e Análise de Dados Genômicos*. 2005. Dissertação (Mestrado em Programa em Pós-Graduação em Informática) - Núcleo de Computação Eletrônica - UFRJ, Rio de Janeiro, 2005.
- Miller GA. WordNet: An on-line lexical database. *Special issue of the International Journal of Lexicography*. 1990; 3(4).
- Mougin F, Burgun A, Bodenreider O. Mapping data elements to terminological resources for integrating biomedical data sources. *BMC Bioinformatics*. 2006; 24(7) Suppl 3:S6.
- Neville HH. Feasibility study of a scheme for reconciling thesauri covering a common subject. *Journal Doc*. dec. 1970 ; 4(26) :313-36.
- Neville HH. Thesaurus reconciliation. *Aslib Proc*. nov. 1972; 11(24): 620-6.
- Noy NF, Musen MA. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000, p. 450-455.
- OBO. *Open Biomedical Ontologies*, 2005. Disponível em: <<http://obo.sourceforge.net>>. Acesso em: 17 maio 2008.
- OWL - *Web Ontology Language*, 2008. Disponível em: <<http://www.w3.org/TR/owl-ref/>>. Acesso em: 17 maio 2008.

Pinto S, Martins JP. A Methodology for Ontology Integration. Proceedings of First International Conference on Knowledge Capture, K-CAP 2001, Victoria, B.C., Canada, ACM Press, 2001.

Ranganathan SR. Prolegomena to Library Classification. New York: Asia Publishing House, 1967.

Reynaud C, Safar B. Exploiting WordNet as Background Knowledge. In: International ISWC'07 Ontology Matching (OM-07) Workshop, Busan, Corea, 2007.

Sabou M, D'aquin M, Motta E. Using the semantic web as background knowledge for ontology mapping. In: 1st International Workshop on Ontology Matching (OM-2006) at ISWC-2006, Athens, Georgia (USA), nov. 2006.

Sales LF. Ontologias de domínio: estudo das relações conceituais e sua aplicação. 141f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense, Rio de Janeiro, 2006.

Silva VS. Alinhamento de ontologias através do algoritmo de Alinhamento Local de Caminhos. Projeto Final de Graduação em Informática – Universidade Federal do Rio de Janeiro, Instituto de Matemática. 2008.

Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. In: Hájek Petr, Valdés-Villanueva Luis, Westerståhl Dag (eds.): Logic, Methodology and Philosophy of Science. Proceedings

of the 12th International Conference, King's College Publications, London, 2005, p. 505-520.

Soergel D. Compatibility of vocabularies. In: RIGGS, F.W. ed. The conta Conference; Proceedings of conference on conceptual and terminological analysis in the social sciences. Bielefeld, may 24-7, 1981. Frankfurt, INDEKS Verl., 1982. p. 209-23.


Stumme G, Madche A. FCA-Merge: Bottom-up merging of ontologies. In: 7th Intl. Conf. on Artificial Intelligence (IJCAI '01), Seattle, WA, 2001, p. 225-230.

Su X. Semantic Enrichment for Ontology Mapping. PhD thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491, Trondheim, Norway, 2004.

Swartout W, Tate A. Guest editors' introduction: ontologies. IEEE Intelligent Systems. jan. 1999; 14(1): 18-9.

Vickery BC. Ontologies. J Info Sci, London. 1997; 23(4):227-86.

Wagner G. Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli*. Dissertação (Mestrado em Biologia Celular e Molecular) - Fundação Oswaldo Cruz. 2006.

Weinstein PC. Ontology-Based Metadata: transforming the MARC Legacy. Digital Libraries, Pittsburg. 1998; p. 254-263. 

## About the authors

### *Maria Luiza de Almeida Campos*

Has a doctorate degree in Information Science from the Brazilian Institute of Scientific and Technological Information - IBICT / UFRJ. Ms. Almeida Campos also has a post-doctorate degree from the Molecular Biology Laboratory of Trypanosomatides and Phlebotomides – Oswaldo Cruz Institute – FIOCRUZ, where she carries out research in the field of genomic ontologies. She is an Assistant Professor of the Post-Graduate Program in Information Science – UFF, and her experience includes teaching and researching in the fields of Information Organization and Retrieval, Taxonomy; Ontology, Construction of Thesauri. Ms. Almeida Campos's activities also include teaching as a guest professor in strictu-sensu post graduation courses of UFRJ's Post Graduation in Informatics (2002-2004) and latu-sensu, both in refresher-level courses (Indexation Course, year 1998-2000 / USU; Knowledge Management Course, year 1998 / USU; Thesaurus Course, year 1994 / UFF; Classification Theory Course, year 1990 / UNIRIO) and specialization-level courses (File Planning, Organization and Directing Course – Information Management, years 1996 - 2007). She was a member of the National Commission on Terminological Principles of the Brazilian Association of Technical Norms – ABNT. Ms. Almeida Campos also carries out a research entitled "Ontology Integration: The domain of Bioinformatics and the issues concerning terminology compatibilization", with a productivity scholarship granted by CNPq. She is the coordinator of the "Ontology and Taxonomy – theoretical and methodological aspects" research group. She has been working as an independent consultant for several institutions, in activities involving the elaboration of taxonomies, thesauri and indexing policies, such as FINEP; Casa de Rui Barbosa; FIOCRUZ; SESC; IPHAN; Central Globo de Produções and Petrobras. She is the author of the book entitled "Documentary Languages: theories underlying their elaboration", and of articles published both locally and internationally.

## *Maria Luiza Machado de Campos*

Is a researcher and professor at the Computer Science Department of the Institute of Mathematics of the Federal University of Rio de Janeiro. Ms. Machado de Campos has a Civil Engineering degree from the Federal University of Rio Grande do Sul and a Masters' Degree in Systems and Computer Engineering from COPPE, Federal University of Rio de Janeiro, as well as a PhD in Information Systems from the University of East Anglia, Norwich, England. She is a scholar researcher level 2 at CNPq. Her areas of activity encompass databases, knowledge management, data warehousing, metadata and ontology management, particularly applied to the domains of bioinformatics, oil and emergencies. Ms. Machado de Campos is an active member in the community for computing research, and she has published several articles in local and international papers. She has also given lectures in the field, provided guidance for numerous essays, dissertations for Masters' and Doctors' degree, and coordinated research projects funded by FINEP, CNPq and FAPERJ. Ms. Machado de Campos has been providing consultancy for companies in the implementation of state-of-the-art technologies focusing on information management, integration, and exploration within organizations.