Information Sciences Methodological Aspects Applied to Ontology Reuse Tools: A Study Based on Genomic Annotations in the Domain of Trypanosomatides†

Maria Luiza de Almeida Campos*, Maria Luiza Machado Campos**, Alberto M. R. Dávila***, Hagar Espanha Gomes****, Linair Maria Campos*, and Laura de Lira e Oliveira*

* UFF-GCI-PPGI/UFF, Rua. Tiradentes 148, Ingá, Niterói, RJ, Brasil, <marialuizalmeida@gmail.com>, <linair@cisi.coppe.ufrj.br>, <llira@gbl.com.br>

** UFRJ-PPGI, Athos da Silveira Ramos s/n, Ilha do Fundão, Rio de Janeiro, RJ, Brasil, <mluiza@nce.ufrj.br>

*** FIOCRUZ/IOC, Av. Brasil, 4365, Manguinhos, RJ, Brasil, <davila@ioc.fiocruz.br>

**** Rua. Tiradentes 148, Ingá, Niterói, RJ, Brasil, <hagarespanhagomes@gmail.com>

Maria Luiza de Almeida Campos is Researcher and Professor, Department of Information Science, Graduate Program, Universidade Federal Fluminense. She has a B.S. in documentation and library science from Universidade Federal Fluminense, and a master's and Ph.D. in information science from Universidade Federal do Rio de Janeiro, covenant with Brazilian Institute of Science and Technology. She held a post-doctorate in Laboratório Biologia Molecular da FIOCRUZ in the area of ontologies. Research interests include knowledge organization, models and theories of knowledge representation, terminology, and foundational ontologies.

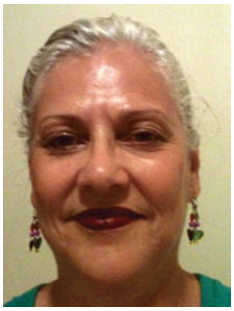
Maria Luiza Machado Campos is Researcher and Professor, Department of Computer Science of the Mathematical Institute of the Federal University of Rio de Janeiro. She graduated in civil engineering from Universidade Federal do Rio Grande do Sul, and has a master's in systems engineering and computation from Coppe, Federal University of Rio de Janeiro and a Ph.D. in information systems from the University of East Anglia, Norwich, England. Her areas of focus include databases, knowledge management, data warehousing, management of metadata, and ontologies, applied particularly to the areas of bioinformatics, oil, and emergencies.

Hagar Espanha Gomes holds a degree in librarianship and documentation from the National Library Foundation, specializing in master's and bibliographic research by the Brazilian Institute of Bibliography and Documentation, and a Doctorate in documentation. Her areas of interest are: classification, terminology, information architecture and representation, and information retrieval.

Alberto Martín Rivera Dávila graduated with a bachelor's in biological sciences from the Federal University of Mato Grosso do Sul (1997) and a Ph.D. in cell and molecular biology from the Oswaldo Cruz Foundation (2002). He is currently a senior researcher of the Instituto Oswaldo Cruz, FIOCRUZ, and he has experience in bioinformatics, computational biology, and molecular biology, mainly in the following areas: bioinformatics and computational biology of protozoa; molecular characterization of protozoa; computational aspects of systems biology; and metagenomics.

Laura de Lira e Oliveira studied at the School of Medicine, Medicine and Surgery of Rio de Janeiro (1974), the State University of Rio de Janeiro (1978) and has a master's of information science (covenant UFRJ / IBICT - 1980), and a Ph.D. in information science (UFF / IBICT - 2011). She has experience in medicine, specializing in cardiology and homeopathy. She has interests in the following topics: classification theory, organization of knowledge representation information, and medical terminology.

Linair Maria Campos is Manager of Information Technology CISI / COPPE / UFRJ, with a master's in computer science at IM / NCE / UFRJ (2004) and Ph.D. in information science from the UFF / IBICT (2011). She has over 25 years of experience in IT, having worked in management, development, and maintenance of information systems. Currently, she is also a substitute teacher at UFF.



De Almeida Campos, Maria Luiza, Machado Campos, Maria Luiza, Dávila, Alberto M. R., Espanha Gomes, Hagar, Campos, Linair Maria, and de Lira e Oliveira, Laura. **Information Sciences Methodological Aspects Applied to Ontology Reuse Tools: A Study Based on Genomic Annotations in the Domain of Trypanosomatides.** *Knowledge Organization. 40(1)*, 50-61. 33 references.

ABSTRACT: Despite the dissemination of modeling languages and tools for representation and construction of ontologies, their underlying methodologies can still be improved. As a consequence, ontology tools can be enhanced accordingly, in order to support users through the ontology construction process. This paper proposes suggestions for ontology tools' improvement based on a case study within the domain of bioinformatics, applying a reuse methodology. Quantitative and qualitative analyses were carried out on a subset of 28 terms of Gene Ontology on a semi-automatic alignment with other biomedical ontologies. As a result, a report is presented containing suggestions for enhancing ontology reuse tools, which is a product derived from difficulties that we had in reusing a set of OBO ontologies. For the reuse process, a set of steps closely related to those of Pinto and Martin's methodology was used. In each step, it was observed that the experiment would have been significantly improved if ontology manipulation tools had provided certain features. Accordingly, problematic aspects in ontology tools are presented and suggestions are made aiming at getting better results in ontology reuse.

Received 29 April 2012; Revised 24 September 2012; Accepted 27 September 2012

† We would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for partially supporting this work.

1.0 Introduction

During the last few years, initiatives of the international scientific community in the field of genomics have led to an explosive growth of biological information, which keeps growing today. The initial concern was the creation and maintenance of databases to store and describe biological data. As genomes continue to be sequenced and described, studies shifted their focus gradually from genome mapping to the analysis of a broad range of information resulting from the functional characterization of genes by means of molecular biology and bioinformatics. In this scenario, it becomes essential to support the interoperation of data obtained through various research projects around the world, interrelating enzymes, genes, chemical components, diseases, cell types, organs, etc. (Mendes 2005).

Ontologies play an essential role in this process, supporting semantic interoperability of heterogeneous distributed systems in a standard way. The Open Biological and Biomedical Ontologies (OBO) Library (OBO 2009) is a terminology repository developed for shared utilization among several biological and medical domains. Among OBO's most disseminated vocabularies, we can highlight Gene Ontology (GO) (Gene Ontology Consortium 2001). GO is a large vocabulary, comprising more than 38,000 terms (http://www.geneontology.org/GO.downloads.ontology.shtml), non-dependent on organism species (Ashburner and Lewis 2002). Still, although GO has a large number of descriptors, other vocabularies are needed in the biomedical domain as we can see by the variety of ontologies available in OBO. It is worth noting that some of those ontologies use several terms that are equivalent to GO terms, and some-

times even contain references to GO terms IDs, as it can be observed in INOH Molecule Role ontology (Yamamoto et al. 2004). This scenario, considering the complexity of building and maintaining such vocabularies, brings about the issue of ontology reuse.

One important aspect of ontology reuse concerns principles adopted for the organization of concepts and their relationships, and also for building definitions associated with such concepts. In this context, this study points towards the importance of investigations within the area of information organization in information science.

Unfortunately, information about such principles is not always available, and, even when it is, vocabularies are built based on different approaches that require conciliation when their reuse is intended. In this context, this study points towards the importance of investigations within the area of language compatibility in information science. Research in this area may provide theoretical and methodological guidelines (Gangemi, Steve, and Giacomelli 1996) that can help make ontology reuse tools more useful and precise. In parallel with the adoption of well founded methodological practices, ontology tools can be improved accordingly to support users throughout the ontology construction process, as well as in providing management strategies for the production and reuse of high quality ontologies.

This paper intends to discuss issues that are inherent to ontology reuse as a methodological step towards acquisition of knowledge in ontologies, and thus propose supporting guidelines for ontology mapping and alignment tools. A case study within the domain of bioinformatics is presented, more specifically focused on genome annotation of trypanosomatides at the BiowebDB consortium (Biowebdb 2006).

This paper is organized as follows: in section 2, common kinds of ontology reuse and related work in computer science is presented; in section 3, we discuss the information science perspective on vocabulary compatibilization; in section 4, some of the issues found in our reuse experience are discussed; in section 5, some semantic aspects of reuse and their impact on ontology tools are presented as result of experiments reusing OBO ontologies; finally, in section 6, future studies are suggested.

2.0 Ontology reuse

Guarino and Musen (2005, 1) highlight the role ontologies have been playing in information systems: "Building ontologies is now an essential activity that underlies nearly everything we do in the development of computational systems." Although Gruber's (1993, 1) is the most commonly cited definition of ontology: "an ontology is the specification of a conceptualization," Guarino (1998, 4) also gives a clear definition:

> In its most prevalent use in AI, an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words.

Ontologies can be reused in many ways depending on users' needs and ontologies' availability. Pinto and Martins (2001) divide reuse processes in "merge" and "integration." In a merging process, a single ontology is created from the reuse (partial or total) of two or more ontologies about the same subject. In an integration process, an adapted (some of the concepts will probably be extended, joined, deleted, or reformulated) and independent ontology is created from the reuse (partial or total) of two or more ontologies on different (although possibly related) subjects.

Some authors (Bruijn et al. 2006; Euzenat and Shvaiko 2007) also include "alignment" as an ontology reuse process. This process, however, differs from those aforementioned by its result; instead of creating an additional ontology, alignment keeps reused ontologies preserved on their original sources, although creating a set of links between terms of the reused ontologies. Such links express the kind of relationship that connects terms from the reused ontologies and are stored in a separate persistent model. This model is the result of a process named term matching (Euzenat and Shvaiko 2007), which aims to identify terms that express similar concepts.

The set of links between ontologies produced by means of the alignment process is a mapping between these 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. One of the issues of mapping concerns how to find candidates. Another aspect involving mapping concerns the type of technique employed to estimate candidates. It can be based, among other aspects: i) on similarities between terms names; ii) on the ontology structure, such as, for in-

stance, considering the terms' positions within the hierarchical structure of ontologies under comparison, or their part-of relations, or even other types of relations (Euzenat and Shvaiko 2007); iii) on the addition of supplementary knowledge, such as information from another ontology or vocabulary with a concept hierarchy, such as Wordnet, which may be used to search for synonyms (Reynaud and Safar 2007).

2.1 Studies related to ontology reuse in computer science

Regarding methodological aspects on how to reuse ontologies, to the best of our knowledge, literature is more often concerned with computational aspects, such as which algorithms are most effective to promote compatibility among ontologies regarding both the accuracy and the speed of their results (Choi, Song, and Han 2006). Nevertheless, some authors propose general tasks that are necessary in the reuse process. Gangemi, Steve, and Giancomelli (2006), for instance, state that it is necessary to identify the basic terms and their necessary and sufficient conditions in textual format. However, they provide no suggestion on how to perform such identification, or on which principles should be used to build the definitions. The more comprehensive view of Pinto and Martins (2001), on the other hand, suggests that the reuse process actually starts during the selection of ontologies to be reused. No systematic details are given though, on how to perform such tasks.

Some of the many studies carried out by Guarino (1998), Barry Smith (2005), and Guizzardi et al. (2011), although not directly focused on reuse per se, may help the process, since they explore the semantic and formal nature of concepts of an ontology. In practice, Guarino's Formal Ontology, as well as Guizzardi's UFO Ontology, can be defined as theories of prior distinctions concerning worldly entities of the world (physical objects, events, regions, amounts of matter); and meta-level categories to model the world (concepts, properties, qualities, states, roles, and parts). Guarino accepts the creation of several, not necessarily complementary, views of a same domain, which he calls "possible worlds." Barry Smith (2005), 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. Smith, as opposed to Guarino and Guizzardi, advocates the idea that there is only one, commonly agreed, "possible world," albeit with different, orthogonal, complementary views.

3.0 Vocabulary compatibilization in information science

Semantic issues have been objects of study and research in Information Science since the beginning of the second half of the last century within a computer environment. Such studies focused construction and compatibilization of documentary languages and their contributions are still valid for compatibilization among and reuse of ontologies.

Two methods distinctly stand out among others used for converting and creating compatibility between languages based on the integration of vocabularies. These are Neville's thesaurus reconciliation method (Neville 1972) and Dahlberg's concept correlation matrix (Dahlberg 1983a). 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 and a series of 11 scenarios with rules to treat vocabulary compatibility issues, which enables the establishment of a conceptual equivalence of descriptors of different languages. The method suggested by Dahlberg is based on the construction of a concept compatibility matrix and a concept register. The concept compatibility matrix provides the results of the language compatibility analysis from the semantic and structural points of view. The first step to elaborate the matrix is the verbal matching of terms. In the second step, additional information supports the understanding of the terms intended meaning by means of a conceptual analysis, whose result is recorded in a concept register. The concept register may be implemented as a database table, although Dahlberg did not propose a solution to implement it computationally. It contains some useful information that helps to identify the semantics associated to each concept, such as: i) the name of the concept in other vocabularies; ii) the concept's form category, which indicates its nature, e.g., if it is an object, a process, a quality; iii) additional information about the concept, for instance, its source; and, iv) related concepts. Recent studies include these issues within KOS (Zeng and Chan 2004). Nevertheless, this paper aims at pointing to a better concept description so that automatic compatibilization procedures work with better precision.

3.1 Information organization in information science

Literature on information organization in the field of information science proved to be helpful, specifically those theories strongly related to representation of concept systems. In those, there are solid European theoretical foundations for the elaboration of documentary languages, providing a semantic base for integration. Examples are: Ranganathan's faceted classification theory (1967) and Dahlberg's concept theory (1978), which allow the representation of knowledge domains. Ranganathan elaborated a series of principles and canons for knowledge classification, which intended to allow concepts of a knowledge domain to be structured in a systematic way. That is, concepts are organized in arrays and chains, which are, in turn, structured in comprehensive classes, called facets, and the latter are organized within a given Fundamental Category. The grouping of all categories comprises a concept system for a given subject area, and each concept within the category is also the manifestation of that category (Ranganathan 1967).

4.0 Case study methodology and results

The purpose of this paper is not a proposal of a reuse methodology, but to show that compatibilization criteria developed in information science are valid to ontology reuse. We may also take advantage of an existing methodology, such as Pinto and Martins (2001), to illustrate how ontology tools can benefit from a joint approach between theory and practice, in the scope of a reuse scenario.

The sample of concepts (knowledge capture) consists of a set of GO terms used in a manual genomic annotation of *Tripanossoma rangeli* made by biologists of the BiowebDB group during a master research project (Wagner 2006). This group of terms constitute a coherent set present in the three branches of GO (cell component, molecular function, biological process); biologists are familiar with those concepts and relations and this is important to validate the structure of these terms when comparing with other ontologies. The result of annotation of *T. rangeli* consists of 865 terms class distributed in those categories.

Five steps of ontology reuse can be summarized: i) finding and selection of candidate ontologies; ii) evaluation of candidate ontologies by domain experts and ontology engineers; iii) final selection of ontologies to be integrated; and, iv) application of operations towards ontology integration, which we consider as a semi-automatic procedure.

4.1.1 Step I: finding and selecting candidate ontologies

To begin with, GO was considered the master ontology. One of the criteria for selecting a master vocabulary is its completeness (Dahlberg 1981). In relation to OBO ontologies, specially, for functional genomic annotation, GO is the most complete and used. So, since the beginning, it was assumed that GO could be considered the master ontology for the experiment.

To identify themes for the compatibilization a domain study was conducted. Many researchers have studied how to approach a given knowledge domain (Soergel 1982, 1997; Lancaster 1986; Hjørland 2002, 2003, 2004; Broughton et al., 2005; Gnoli and Hjørland 2009). They provide us with systematic guidelines for a preliminary domain analysis. Support provided by these theoretical contributions and by others from the social sciences (Latour 1997) have allowed elaboration of a preliminary draft of thematic groupings on the domain of trypanosomatides. At first, ten thematic groups were identified: protists; functional and systems biology; molecular biology and genomics; evolutive molecular genetics; comparative genomics; philogeny; bioinformatics; diseases; and metagenomics; targets for drugs, each one with its own sub-groupings.

The purpose of this selection was to identify a set of ontologies to be reused with the aid of software tools. This strategy is in accordance with Neville's feasibility study for reconciliation of thesaurus (Neville 1972) and with Dahlberg's intermediate language proposal for compatibilization (Dahlberg 1983b). The intermediate language—or 'master' vocabulary—would be the starting point when establishing equivalence relations with terms of other ontologies.

To identify possible useful ontologies for the experiment a search was made in the OBO site, where each ontology has a brief summary of its scope. Through this, it was possible to identify those in accordance with the thematic areas previously chosen and, using this opportunity, verify their ontological commitment. This scope analysis showed, for example, that 'molecular role' in one ontology does not refer to molecular role, but is, indeed, an ontology of proteins (Campos 2011). The result of this step led to a selection of eleven ontologies that could be of interest to researchers on trypanosomatides.

4.1.2 Step II: evaluation of candidate ontologies by domain experts and ontology engineers

Selection of candidate ontologies for the experiment was validated through seminars with the research

group. To support the ontology evaluation process, Onto-Edit was used, as it allows user-friendly visualization of concepts and hierarchies. Such visualization shows clearly taxonomies inherent to each ontology and is useful for compatibilization and, the case being, for further integration. (Jie, Fei, and Sheng-Uei 2011) From the initial group of ten ontologies, six were confirmed by end-users as being of interest. These had already been used as knowledge source so that classes of interest within the domain of Trypanosomatides could be easily identified.

4.1.3 Step III: selecting ontologies

Laboratory researchers selected the following ontologies: NCBI organismal classification, pathway, sequence types and features (SO), Brenda tissue/enzyme source, Event-INOH pathway ontology, multiple alignment and system biology (Open Biomedical Ontologies 2009).

4.1.4 Step IV: applying operations towards ontology integration

Two procedures were required in order to evaluate the degree of compatibilization between GO and each selected ontology: the identification of hierarchies among selected ontologies, and the semantic analysis of each term within each hierarchy. These also contributed to verifying their potential reuse.

4.2 Identification of hierarchies

Visualization of hierarchies was done through application of the OBO-Edit editor (Day-Richter, Harris, and Haendel 2007), which supports multiple visualization forms. Other requirements not supported by this tool were needed, such as facilities for recording justification of the choices made and thematic superimposition.

Mapping terms in selected ontologies was done with application of Prompt tool. To allow more effective searching and exploring the hierarchy of terms in several biological ontologies with dynamic trees, with ontology subsets and retrieving of information, OntoExplore was developed. This tool was used afterwards in the process of genomic annotation thus providing researchers with a computational support.

OntoExplore was developed in Java, using the API JENA (http://jena.sourceforge.net) to parse ontologies in OWL and RDF formats and the Prefuse Visualization Toolkit (http://prefuse.org) to implement interactive data visualization mechanisms.

OntoExplore allows: (i) Visualization and comparison of terms hierarchies in different ontologies. It is possible to select a term and visualize its hierarchy in two different ontologies (see Figure 1). Thus it is possible to check and study the hierarchy of terms. It also allows (ii) the searching of terms within multiple ontologies. The goal is to find similar terms. Sometimes the term exists in another ontology with a different name. To implement this, a synonym-based search was applied.

The purpose of OntoExplore is to align ontologies by means of an algorithm that explores their hierarchical structure, the term, and also the semantic nature of the concepts, according to the Classification Theory (Ranganathan 1967). For the latter to be possible, the root classes of two of the reused ontologies were previously manually associated to terms denoting Fundamental Categories.

Our goal in building such tool instead of using existing ones, like Prompt (Noy and Musen 2003), is to implement and test some of the aspects we consider important to ontology mapping in order to evaluate its helpfulness on the reuse process. The use of Fundamental Categories is an example of such aspects.

From 865 GO terms, 28 were common in selected ontologies. That means that each term was listed at least once in each ontology besides GO. In great measure, terms were found only in GO and Event (INOH pathway ontology). This result suggests that this ontology has enough thematic superimposition with GO. It deals with biological events such as mechanisms of gene expression and immunological response, concepts that belong to the Functional Genomics domain, and biological process is one of the three components of GO.

This paper limits discussion to results of experiments between GO and INOH.

4.3 Semantic analysis

Once a term is found in a target ontology, a subset is derived from such an ontology, composed of its ascending and descending hierarchy within GO and INOH. Mapping is done with the assistance of the Prompt tool and with our prototypical tool. Each resulting mapping is then manually analyzed based on: (i) similarities in term designations; (ii) semantic similarity indicating concepts of similar nature (logically related); (iii) relations indicating concepts that are not similar, but that may be associated by means of category (logic) relations which are relevant to the domain, for instance, between a protein and a bio-

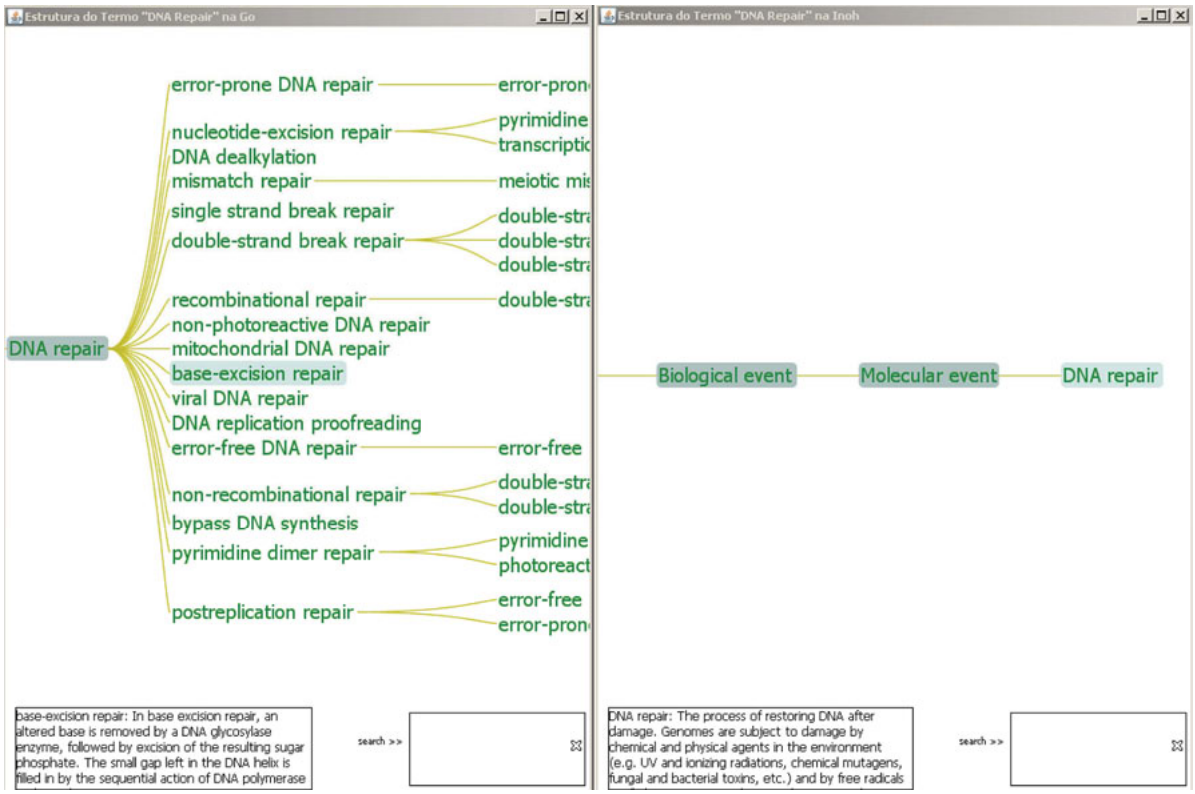


Figure 1. Comparing the same term hierarchy in two ontologies

logical process in which it participates, a biological process and its product.

Term definitions were also analysed according to methodological principles (Dahlberg 1978a, 1978b, 1981, 1983a; 1983b; Neville 1972). A manual analysis and comparison of terms and their definitions both in GO and INOH were made, aiming at verifying how much future automatic processing would provide consistent results relating to semantic aspects. The analysis made measuring the degree of verbal and conceptual compatibility possible. Verbal coincidence was the starting point, following analyses of definitions to obtain conceptual coincidence. When no definition was provided, it was necessary to observe term position in its respective hierarchy so that similarity of classification between ontologies could be observed. Manual analysis of definitions was made to verify the degree of consistent results when applying future automatic processing; in other words, to evaluate semantic potentiality of compatibilization when comparing common characteristics between definitions.

Dahlberg's matrix of semantic compatibility starts with verbal coincidence. The **rate of verbal coincidence** indicates the possibility of measuring the degree of conceptual compatibility. Two measures were investigated:

– Concept coincidence
– Concept correspondence.

Homonyms were also investigated.

4.3.1 Conceptual coincidence

Conceptual coincidence occurs when for the same verbal form and same content 80% of characteristics occur in both definitions. In this case, 31% of terms are considered conceptually identical. But two different situations were identified in relation to hierarchical structure:

1 – Some possess the same generic term: in both ontologies, "cell-cell signaling" is subordinated to "cell communication." It is worth observing that cell-cell signaling definition in both ontologies is: "Any process that mediates the transfer of information from one cell to another."

2 – Terms have different generic term: "DNA repair"—in GO, it is subordinated to "DNA metabolic process"; in INOH, it is subordinated to "molecular event." It is worth noting that "DNA Repair," in both ontologies, is defined as: "The process of restoring DNA after damage. Ge-

nomes are subject to damage by chemical and physical agents in the environment (e.g., UV and ionizing radiations, chemical mutagens, fungal and bacterial toxins, etc.) and by free radicals or alkylating agents endogenously generated in metabolism. DNA is also damaged because of errors during its replication. A variety of different DNA repair pathways has been reported that include direct reversal, base excision repair, nucleotide excision repair, photoreactivation, bypass, double-strand break repair pathway, and mismatch repair pathway."

The analysis of these terms indicates that, although they seem conceptually identical, according to their definitions, they have different hierarchies, and a conflict will rise in an automatic analysis when determining conceptual similarity. This is due to lack of a definition pattern that ensures that the first element in definition be its immediate superordinated term in the conceptual chain. In this case, they were considered identical.

4.3.2 Conceptual correspondence

Conceptual correspondence occurs when the same verbal form and similar concept content are considered quasi-synonyms when 60-79 % common characteristics occur in both definition. In this case, 63% of terms can be considered quasi-synonyms. "Organ morphgenesis" is an example.

In INOH, the term has the following definition: "Morphogenesis of a tissue or tissues that function together to perform a specific function. Organs are commonly observed as visibly distinct structures, but may also exist as loosely associated clusters of cells that function together as to perform a specific function." In GO, the definition is as follows: "Morphogenesis of an organ. An organ is defined as a tissue or set of tissues that work together to perform a specific function or functions. Morphogenesis is the process by which anatomical structures are generated and organized. Organs are commonly observed as visibly distinct structures, but may also exist as loosely associated clusters of cells that work together to perform a specific function or functions."

4.3.3 Homonyms

Only two terms (6%) in each ontology were semantically different so they were considered homonyms (for example, "Phosphorylation"). The definition in INOH is as follows: "Reversible reaction that can affect D,C,H,S,T,Y,R residues." The definition in GO is: "The process of introducing a phosphate group into a molecule, usually with the formation of a phosphoric ester, a phosphoric anhydride or a phosphoric amide."

Analyses showed that, due to lack of a pattern, definitions do not allow consistent results in an automatic processing for semantic compatibility degree between concepts. As it could be observed, besides having conceptual coincidence (identical definitions), two identical terms cannot be considered identical because each hierarchical structure does not match.

To obtain consistent semantic compatibilization degree between ontologies, interference in definitions will be needed; or, to provide quantitative analyses of similar characteristics as well as to be able to verify superordinated term in each hierarchy, software will have to be developed. The correspondence between concepts would be better obtained if granularity, synonymy, and establishment of principles for a standard terminology were previously established.

The use of categories associated to ontologies was one of the functionalities aggregated to OntoExplore. It resulted in an increased accuracy when handling false positives, which brings us closer to the ideal set of intended mappings. As an example, we can mention the case 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 both ontologies are mapped through the Prompt tool, it indicates that the terms are similar but they actually require a semantic analysis. Similarly, the terms "transporter," from the MoleculeRole 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." As it can be seen, these term pairs, despite their linguistic similarity, denote concepts with distinct natures (different categories), therefore Prompt should not have suggested those terms as mapping candidates. A person can observe this but the tool provided no mechanisms to register it, so one will deal with this same issue when trying to align ontologies.

In this context, it was possible to manually confirm a suggested 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, "transporter" (a protein, in MoleculeRole) and "participates in transport" (a process, in GO), suggests a relation between a biological object (a protein) and a process (transport). In this scenario, OntoExplore provides mechanisms, absent in Prompt, to persist in the acknowledgement of the validity of such a relation, so an alignment attempt could be made in an incremental and more precise way.

5.0 Semantic aspects of reuse and the impact on ontology tools

During this work on OBO ontologies, several aspects of importance to ontology reuse have been found, considering not only machines, but also humans, such as: (a) concept comprehension; (b) concept categorization; (c) concept definition; (d) ontological commitment elucidation; (e) concept matching; and (f) ontology articulation.

Aspect (a) regards showing people (and not machines) information regarding the ontology as a whole (for instance, its purpose and design rationale) and about the intended meaning of each term on each ontology as accurately as possible. This may be helpful when people are trying to understand the perspectives used by different ontologies to represent domain knowledge.

Aspect (b) regards providing people some input about the principles by which ontology categories are organized. This may be particularly helpful if one intends to extend the ontology or relate it with another one, because it aids preventing ambiguous categorization or association. Some of these principles can be formalized in order to be used by tools, for instance, in the context of ontology alignment.

Aspect (c) regards improving consistency among the definitions of terms and, with such well-formed definitions, help people to organize and extend ontology taxonomy structure. Besides, the use of standard definitions can improve the results of ontology tools (for instance in mining operations, to propose relations between terms), which can be configured to take advantaged of such semi-structured information.

Aspect (d) regards helping people to evaluate and decide if the ontologies considered in a first selection are useful to the purpose they have in mind.

Aspect (e) regards providing an overview of issues encountered during an ontology compatibility enterprise. Although it may be difficult to keep such records up to date when ontologies change, it is worth keeping this information available to an organized ontology community (such as OBO) as a feedback of a process of ontology compatibilization; it can be used to improve ontologies evolution.

Finally, aspect (f) regards helping users envision possibilities of extending the scope of a particular ontology by connecting terms on this ontology with terms of another ontology that may complement it.

It is worth noting that this list does not pretend to be exhaustive, but, instead, is a proposal of a set of issues derived from our own experience reusing a set of OBO ontologies. Besides, we have observed that the aspects presented on our list are present among the steps realized in a reuse process, and so, accordingly, should be present somehow on an ontology reuse methodology.

It is assumed that a more consistent and accurate reuse can be achieved if ontology tools reflect the multiple aspects and steps of ontology reuse accordingly. Some of those aspects, following the theories presented so far and the experiments conducted in the Biowebdb project, are illustrated in Table 1. Their usage, as situated within the steps of a reuse methodology, may improve the precision of ontology compatibilization mechanisms. This happens because they enhance the semantics associated to ontologies concepts and help users to accomplish most of the tasks carried out on each step of such a reuse methodology.

6.0 Conclusion

Although the aforementioned tools provided valuable help on reusing ontologies, especially regarding the task of finding candidate terms (matching) for mapping, many of the features observed as useful are lacking while reusing ontologies. In this sense, further studies will investigate whether the application of the proposed suggestions, based on semantic aspects of reuse, contribute to an increase in the accuracy using software tools. Future enhancements, modifications and investigations are necessary to improve Onto Explore such as to provide a broad set of metrics to compare term hierarchies in distinct ontologies.

This paper points to the need of systematization of definitions when constructing semantic tools. It is important to follow a pattern that reveals the nature of concepts and their epistemological contexts, or the

Aspect	Tools should tackle	How to do it
(a) Concepts comprehension	Exhibit the matching concepts definitions alongside with their main hierarchic structure.	Multiple ontology visualization mechanisms, showing terms' definitions and other relevant information, such as concepts' categorization and ontological commitment.
(b) Concepts categorization	Support documenting and viewing the fundamental categories under which the ontology has been built.	Standard metadata associated to each concept of each ontology, possibly stored as a concept register, similar to Dahlberg's proposal, on an ontology repository provided along by the tool.
(c) Concepts definition	Allow definition of concepts based on patterns, possibly associated to underlying fundamental categories identified.	Standard metadata associated to each category, suggesting attributes that should be present on the definition of a concept belonging to such category.
(d) Ontological commitment elucidation	Allow documenting and viewing the principles under which the ontology has been built, its purpose, scope, subject, and premises, among others.	Standard metadata associated to each ontology, possibly stored on an ontology repository provided along by the tool.
(e) Concepts matching analysis	Show compatibility issues, as presented by Neville, e.g. difference in granularity; different number of terms to denote the same concept; synonyms; homonyms; Provide support to suggest standard term names;	Standard metadata, related to such compatibility issues, associated to the equivalence relationship between matched concepts.
(f) Ontologies articulation possibilities	Offer suggestion of relationships between concepts on different ontologies, possibly underpinned by categorical relations occurring between concepts that belong to those categories.	Through analysis of concepts definition, and domain analysis; e.g., when one concept refers to another already existent in one of the ontologies to be articulated and the concepts involved belong to categories that may be related.

Table 1. Suggestions to improve precision on ontology reuse tools

best computing tools will continue to produce unsatisfactory results.

The biomedical domain is complex and challenging. In this scenario, the experiment points towards suggestions for ontology tools improvement. Existing tools lack mechanisms to deal accurately with large and multiple ontologies to help users understand their purpose, subject, scope, and ontological commitment.

This paper proposes enhancements that can be performed by ontology tools in order to provide features consonant with ontology reuse methodologies. Such enhancements, if existent, would have been of great utility, as pointed out in the experiment.

The experiment suggests the possibility of applying theoretical principles of compatibilization of documentary languages to ontology domain aiming at obtaining a better classification in a taxonomy.

References

Ashburner, Michael and Lewis, Suzanna. 2002. On ontologies for biologists: the gene ontology – uncoupling the web. In Bock, Gregory and Goode, Jamie A., eds., *'In silico' simulation of biological processes: novartis foundation symposium 247*. Chichester, UK: John Wiley & Sons, pp. 66-80.

BiowebDB Consortium. [2009]. *Comparative genomics approaches*. Available http://biowebdb.org/.

Broughton, Vanda; Hamsson, Joacim; Hjørland, Birger and López-Huerta, Maria José. 2005. Knowledge organization. In: Kayberg, Leif and Lørring, Leif, eds. *European curriculum reflections on Library and Information Science education*. Copenhagen: Royal School of Library and Information Science, pp. 133-48.

Bruijn, Jos, Ehrig, Marc, Feier, Cristina, Martíns-Recuerda, Francisco, Scharffe, François and Weiten, Moritz. 2006. Ontology mediation, merging and aligning. In Davies, John, Studer, Rudi and Warren, Paul, eds., *Semantic web technologies: trends and research in ontology-based systems*, Chichester, UK: John Wiley & Sons.

Campos, Linair M. 2011. Diretrizes para definição de recorte de domínio no reúso de ontologias biomédicas: uma abordagem interdisciplinar baseada na análise do compromisso ontológico. PhD dissertation. Universidade Federal Fluminense / Instituto Brasileiro de Informação em Ciência e Tecnologia

Choi, Namyoun, Song, Il-Yeol and Han, Hyoil. 2006. A survey on ontology mapping. *SIGMOD record* 35 no. 3: 34-41.

Dahlberg, Ingetraut. 1978a. A referent-oriented, analytical concept theory of INTERCONCEPT. *International classification* 5: 142-51.

Dahlberg, Ingetraut. 1978b. Teoria do conceito. *Ciência da informação* 7: 101-07.

Dahlberg, Ingetraut. 1981. Towards establishment of compatibility between indexing languages. *International classification* 8: 88-91.

Dahlberg, Ingetraut. 1983a. Conceptual compatibility of ordering systems. *International classification* 10: 5-8.

Dahlberg, Ingetraut. 1983b. Terminological definitions: characteristics and demands. In *Problèmes de la définition et de la synonymie en terminologie*. Québec: Girsterm, 13-51.

Day-Richter, John, Harris, Midori A, Haendel, Melissa, The Gene Ontology OBO-Edit Working Group and Lewis, Suzan. 2007. OBO-Edit-an ontology editor for biologists. *Bioinformatics* 23: 2198-200.

Euzenat, Jérôme and Shvaiko, Pavel. 2007. *Ontology matching*. Berlin: Springer Verlag.

Gangemi, Aldo, Steve, Geri and Giacomelli, Fabrizio. 1996. ONIONS: an ontological methodology for taxonomic knowledge integration. In *ECAI-96 workshop on ontological engineering*.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. *Genome research* 11: 1425-33.

Gnoli, Claudio and Hjørland Birger. 2009, Letter to the editor: Phylogenetic classification revisited. *Knowledge organization* 36: 78-79.

Gruber, Thomas. R. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition* 5: 199-220.

Guarino, Nicola. 1998. Formal ontology in information systems. *Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998*. Amsterdam: IOS Press, pp. 3-15.

Guarino, Nicola and Musen, Mark A. 2005. Applied ontology: focusing on content. *Applied ontology* 1: 1-5.

Guizzardi, Giancarlo, Almeida, João Paulo, Guizzardi, Renata S.S., Barcellos, Monalessa P. and Falbo, Ricardo. 2011. Foundational ontologies, conceptual modeling and semantic interoperability. *Proceedings of the Iberoamerican meeting on ontological research*. Available http://iaoa.org/isc2012/docs/Guarino2005_Focusing_on_content.pdf.

Hjørland, Birger. 2002. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of documentation* 58: 422-62.

Hjørland, Birger. 2003. Fundamentals of knowledge organization. *Knowledge organization* 30: 87–111.

Hjørland, Birger. 2004. Arguments for philosophical realism in library and information science. *Library trends* 52 no. 3: 488–506.

Jie, Xie, Fei, Liu and Sheng-Uei, Guan. 2001. Tree structure based ontology integration. *Journal of information science* 37. 594-613.

Lancaster, Frederick W. 1986. *Vocabulary control for information retrieval*. 2nd ed. Arlington, VA: Information Resources Press.

Latour, Bruno. 1997. *Ciência em ação: como seguir cientistas e engenheiros sociedade afora*. São Paulo: Editora Unesp.

Mendes, Pablo N. 2005. *Uma abordagem para construção e uso no suporte à integração e análise de dados genômicos*. M.A. thesis. Federal University of Rio de Janeiro.

Neville, Hugh Henry. 1972. Thesaurus reconciliation. *Aslib proceedings*., 24: 620-6.

Noy, Natasha. F. and Musen, Mark, A. 2003. The PROMPT suite: interactive tools for ontology merging and mapping. *International journal of human-computer studies* 59: 983-1024.

Open Biomedical Ontologies. 2009. Available at: http://www.obofoundry.org.

Pinto, Helena Sofia and Martins, João P. 2001. A methodology for ontology integration. *K-CAP '01 proceedings of the 1st international conference on knowledge capture*: 131-8.

Ranganathan, Shiyali R. 1967. *Prolegomena to library classification*. New York: Asia Publishouse.

Reynaud, Chantal and Safar, Brigitte. 2007. Exploiting WordNet as background knowledge. In *International ISWC'07 ontology matching (OM-07) workshop, Busan, Corea*.

Smith, Barry. 2005. The logic of biological classification and the foundations of biomedical ontology. In *Logic, methodology and philosophy of science. Proceedings of the 12th international conference*, *London*, pp. 505-20.

Soergel, Dagobert. 1982. Compatibility of vocabularies. In *Proceedings of conference on conceptual and terminological anlysis in the social sciences*. *Bielefeld, Frankfurt*, pp. 209-23.

Soergel, Dagobert. 1997. Multilingual thesauri and ontologies in cross-language retrieval. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval Stanford University, March 24-26, 1997*. Available http://www.dsoergel.com/cv/B60 .pdf.

Wagner, Glauber. 2006. Geração e análise comparativa de seqüências genômicas de Trypanosoma rangeli. M.A. thesis. Instituto Oswaldo Cruz.

Yamamoto, Satoko, Asanuma, Takao, Takagi, Toshihisa and Fukuda, Ken I. 2004. The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. *Comparative and functional genomics* 5: 528-36.

Zeng, Marcia L. and Chan, Lois M. 2004. Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology* 55: 377-95.