Extracting new Relations to Improve Ontology Reuse

Miguel G. P. de Carvalho[1], Linair M. Campos[2], Vanessa Braganholo[3],
Maria Luiza M. Campos[1], Maria Luiza A. Campos[4]

[1] Universidade Federal do Rio de Janeiro – PPGI/UFRJ, Brazil
{miguelgabriel, mluiza}@ufrj.br
[2] Universidade Federal do Rio de Janeiro – CISI/COPPE/UFRJ, Brazil
linair@cisi.coppe.ufrj.br
[3] Universidade Federal Fluminense – IC/UFF, Brazil
vanessa@ic.uff.br
[4] Universidade Federal Fluminense – PPGCI/UFF, Brazil
maria.almeida@pq.cnpq.br

Abstract. Ontologies reuse in biomedicine faces several challenges, which include the complexity of the domain and the articulation of different vocabularies, many of them with thousands of terms. Those ontologies are still subject to changes to improve their quality and to solve existing deficiencies. Among them, we highlight the problem of concepts that should be related, according to their definition, but are not explicitly connected through relations in the ontology. This could hinder ontology comprehension and limit the scope of the domain represented by the vocabulary, which can eventually encourage the development of new ontologies instead of reusing existing ones. Due to these issues, the adoption of ontology tools to support the discovery of implicit relations, intra and inter ontologies, is strongly recommended. Although there are tools geared to find new relationships in ontologies, such tools fail to consider the specific rationale used to organize the knowledge related to a specific domain. This paper proposes an approach for information extraction in ontologies, which uses the definition and nomenclature of each concept to extract implicit information that can complement the knowledge they contain. This increase in knowledge allows an increase in the quality and semantics of ontologies, and thus improves several processes, including the reuse. We have applied the approach to the biomedical domain and, as a result, we have discovered a set of possible new relations for a single ontology as well as relations between two different ontologies.

Categories and Subject Descriptors: I.2 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*

Keywords: aligment, information extraction, ontology

1. INTRODUCTION

In the last years, with the advance of information and communication technologies, scientific information has become more dynamic, distributed and complex. In this scenario, the use of ontologies is of uttermost importance, as they allow the formal representation of a knowledge domain, allowing the understanding of knowledge by humans and computational treatment in a more efficient way, by machines.

In the biomedical domain (which aggregates the areas of Medicine and Biology), ontologies have become an important tool aiding the description (annotation) of experiments, especially the ones related to genomics, which involve a multitude of diverse concepts, such as those related to biological processes, cellular component, molecular and gene functions. Genome annotation is a process where

We would like to thank the Biowebdb Consortium, coordinated by Dr. Alberto Dávila, at Oswaldo Cruz Institute, Rio de Janeiro, for the insightful discussions and for providing us the data for our experiments. We also would like to thank CNPq and FAPERJ for partially supporting this work. Maria Luiza M. Campos, Maria Luiza A. Campos and Vanessa Braganholo are sponsored by CNPq. Vanessa Braganholo and Maria Luiza M. Campos are also supported by FAPERJ.

a researcher, sometimes called annotator, describes genes with the means of standard vocabularies and referential sources. Standard terms from previously annotated genomes can be used, by analogy, as basis for describing other genomes that contain similar sequences of nucleotides. The most used ontology for this purpose is Gene Ontology[1], although other ontologies from OBO Consortium (Open Biological and Biomedical Ontologies Consortium)[2] have also been used [Silva 2010].

The OBO Consortium is a collaborative consortium involving several groups of developers of biomedical and biological ontologies. The objective of this consortium is defining patterns and guidelines for creation, maintenance and evolution of ontologies in this area, in order to allow them to be interoperable and shared by several biological and medical domains. Their efforts are focused on enhancing the scope, documentation and quality of the ontologies [Smith et al. 2007].

Although the OBO manifest tackles the issue of avoiding the redundancy of descriptors by means of orthogonal subjects, this led to the fragmentation of biomedical knowledge among several ontologies (some more specialized and others more general). Such fragmentation, in turn, can hinder the global understanding of the domain. In this scenario, ontology reuse, especially by means of establishing equivalence (alignment) between concepts of ontologies on related subjects, can substantially contribute to the comprehension of the domain. This is due to the fact that alignment allows the combined use of different ontologies, preventing the annotator to separately search for terms in each of the ontologies of interest. In this sense, by means of ontology reuse, it is possible to improve, among other things, the process of genome annotation, which can make use of additional knowledge to underpin the choice of the most adequate term for annotation and, possibly, enhance the quality of the resulting annotation [Silva 2010].

However, biomedical ontologies face several problems which can hamper their reuse. These problems are materialized into badly documented vocabularies, some containing thousands of terms [Smith et al. 2003]. Many of these vocabularies are built based on ill-formed hierarchies, which pose obstacles to their use and understanding. But, being around for so long, they are nowadays widely used as *de facto* standards in annotations [Smith and Kumar 2004], which hardens their reformulation.

Many of the problems found on biomedical ontologies have already been reported in the literature [Smith et al. 2003; Smith and Kumar 2004; Wroe et al. 2003; Smith et al. 2004; Smith and Rosse 2004; Ogren et al. 2004; Ogren et al. 2005; Smith et al. 2005; Kohler et al. 2006; Schulz et al. 2006; Gu et al. 2009], and reveal aspects from different natures, such as:

(i) problems in structure: lack of relevant terms; lack of coherence in the structure of concepts (very general concepts directly connected to very specific concepts);

(ii) problems in relations: few relations are available to express domain knowledge; overload of relations; omission of apparently obvious relationships;

(iii) problems in terms definitions: implicit information contained in definitions are inconsistent with explicit knowledge represented in the ontology; circular definition;

(iv) contextual problems: lack of formal documentation; lack of descriptors on the theme and scope of the ontology; lack of more formal patterns for OBO ontologies construction (causing, for instance, homonymy and synonymy); lack of sound ontology documentation, presenting their scope, objective and theme, which sometimes causes overlap or conflict on scope.

These problems may be identified for future correction or to facilitate ontology reuse, by means of ontology information enrichment, i.e., by making explicit knowledge that was previously implicit, improving the expressed semantics and supporting richer inferences. As an example, we can quote the definition associated to the term Mitochondrion (GO:0005739): "A semiautonomous, self replicating

[1] http://www.geneontology.org/

[2] http://www.obofoundry.org/

organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration"[3]. In this definition, it is possible to extract three relations, which are not explicitly found in the ontology:

(i) ***is a*** semiautonomous, self replicating organelle;
(ii) ***occurs in*** varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells;
(iii) ***the site of*** tissue respiration.

Such relations are candidates for future improvement in the knowledge represented in the ontologies. For example, they can be used to validate the consistency of existing relationships and complement the knowledge in the ontology.

Our research is related to ontology enrichment and, in this paper, we present an approach to extract information hidden on names and definitions contained on ontologies, underpinned by four pillars: use of dictionaries, rules, machine learning and statistics. The experiment conducted to validate our approach is focused on the extraction of the hierarchical relation "is-a". We have chosen this relation because ontology hierarchies are usually structured according to it, and so it is directly tied to the ontology reuse process. On applying our approach, we have found the existence of implicit knowledge in biomedical ontologies and we suggest ways to deal with such knowledge in order to improve the ontology reuse process and, hence, to improve the genome annotation processes as well.

The remaining of this paper is organized as follows: in Section 2 we present the ontology typology, the "is-a" relation and the importance of concepts definitions to reveal implicit relationships. Section 3 presents existing strategies to perform information extraction that are relevant to our approach. In Section 4 we discuss related work. In Section 5 we present our approach and in Section 6 we describe an experiment conducted to extract new relationships on biomedical ontologies. Finally, we conclude in Section 7.

2. CONCEPTUAL BACKGROUND

Although ontologies are not always built with the same conceptual metamodel, certain aspects and basic components are usually present in most of them. In general, ontologies are composed by terms, relations and definitions [Sales et al. 2008]. **Terms** are components which denote the concepts in the domain. Concepts are connected to each other through **relations** and are described in more detail by their **definition** [Sales et al. 2008].

In ontologies, the most important relation is "is-a", because this relation is responsible for describing terms hierarchies [Brachman 1983]. The relationship "is-a" is also known as hierarchical relation, generic relation or generalization/specialization [Brachman 1983; Dahlberg 1978; 1981]. It assigns a semantic of types and subtypes when relating concepts (classes or terms) and subconcepts (subclasses or subterms) [Brachman 1983]. This can be related to sets theory, where a broader set (more general concept) contains a more specific set (more specific concept) [Smith et al. 2005].

Altogether with relations, term definition on ontologies contributes to the comprehension of a concept meaning by humans. Several works [Dahlberg 1978; 1981; Michael et al. 2001; Kohler et al. 2006] tackle the importance of a well-structured and formalized term definition, allowing not only comprehension by humans, but also knowledge inference by machines. In this context, problems on concepts definitions have been investigated on several studies along the years. In 1982, for instance, the *Groupe Interdisciplinaire de Recherche Scientifique et Appliquée en Terminologie* (GIRSTERM) discussed the particularities of definitions in Terminology (e.g., What to define? How to define? Why define? Does term synonymy can be compared to word synonymy?) on an event named "International Colloquium

[3] `http://www.geneontology.org/`

on Terminology", under the title "The problems of definition and synonymy in terminology" [Campos 1994; Gomes and Campos 2004].

Aware of such issues, Dahlberg (1978, p.149) proposed definition to be: "the establishment of an equivalence between the term (the *definiendum*) and the necessary characteristics of the referent of a concept (the *definiens*) for the purpose of delimiting the use of the term in discourse". Still according to Dahlberg (1978), the most usually known structure of *definiens*, and also the oldest one, is *genus proximum and differentia specifica*. This structure of *definiens* relates a given concept to a broader one or to an equivalent one (*genus proximum*) and specifies these concepts through a description of their characteristics (*differentia specifica*), thus enabling the main concept to be related to something familiar or close to the observer [Dahlberg 1978; Gomes and Campos 2004]. With certain frequency, we can find the adoption of the definition structure *definiens, genus supremum* in ontologies. These *definiens* usually denote specifying features that designate "property", "category", and thus work as hierarchical relationships, since they relate a given concept to a more general concept (*genus supremum*) [Dahlberg 1978; Gomes and Campos 2004]. In biomedical ontologies the *genus supremum* is observed more frequently than the *genus proximum*, as, for instance, in definitions that begin with "A process cycle ...", "Any process . . . " and "A chemical reaction ...". Such definitions exemplify a hierarchical relation between the described term and the term used in the definition.

The comprehension of terms relations, hierarchies and definitions is of prime importance to an effective information extraction procedure, as the identification of patterns allows for more appropriate means to deal with ontologies. Through this analysis, one can take advantage of exploring such resources aiming at information extraction, enrichment and also in search of coherence within ontologies.

3. INFORMATION EXTRACTION

Information extraction can be defined as an application of natural language processing with the goal of extracting information from a corpus [Ananiadou and McNaught 2006]. Information extraction usually occurs in two steps. The first step uses natural language processing, and, in the second step, a series of techniques and approaches are applied to extract information. Although these steps have both their peculiarities, in many works they may be joined into one single step [Ananiadou and McNaught 2006].

Natural language processing can be defined as the set of approaches, techniques and mechanisms used to analyze text corpora written in natural language. After the corpus is treated by natural language processing, the second step of information extraction has the function to identify and extract information. For this, several approaches can be used. The main strategies for information extraction on the biomedical domain, according to Ananiadou and McNaught (2006), are:

(i) **dictionary-based**: uses resources from terminologies (dictionary) to mark and extract concepts in the corpus.

(ii) **rule-based**: uses several formation patterns to find and extract relevant concepts in the corpus;

(iii) **machine learning-based**: uses machine learning techniques to identify useful concepts in the corpus through previously received training;

(iv) **statistics-based**: uses statistical techniques to calculate different distributions of text inside the corpus in order to identify and extract information.

Table I shows a comparison of these approaches, showing the advantages and disadvantages of each of them.

Table I. Comparison of approaches for information extraction in Biomedical Ontologies

Strategy	Advantages	Disadvantages
Dictionary-based	perfect match with the dictionary	does not identify neologisms; ambiguous terms decrease the accuracy
Rule-based	has better accuracy than other approaches	new rules are needed for each new domain
Machine learning-based	supports inference of information in the corpus	depends on the training received
Statistics-based	can be more easily adapted to different domains than the other approaches	requires definition of appropriate statistical measures

Source: Based on Ananiadou and McNaught (2006)

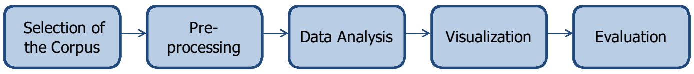


Fig. 1. Five-steps approach for information extraction in biomedical domain, according to Mathiak e Eckstein (2004)

4. RELATED WORK

The identification of relations and interactions among genes, molecular function, chemical reaction, enzymes, drugs, diseases, proteins and other biomedical concepts is important to researches in this area. Considering the volume of information involved, the utilization of text mining and information extraction has become increasingly necessary to aggregate new knowledge to scientific research [Palakal et al. 2005].

Several approaches and mechanisms [Agarwal and Searls 2008; Fundel et al. 2007; Hakenberg et al. 2010; Krallinger et al. 2008; Rinaldi et al. 2004; Rinaldi et al. 2007] developed for the biomedical domain use as corpus their own literature. As an example, several of such approaches use thesis, dissertations and scientific articles available on the Internet, especially those available on specialized bases such as PubMed. Most of these works consider a sequence of extraction activities, similar to the one proposed by [Mathiak and Eckstein 2004]. This approach uses natural language processing in order to prepare the corpus for information extraction per se, including five steps for biological information extraction (Fig. 1). The problem with this kind of approach is that it does not take into consideration the particularities of each corpus. Besides, these approaches usually adopt only one strategy to extract information, most of the time using either rules or dictionaries.

As far as the OBO consortium ontologies are concerned, the OBOL tool [Mungall 2004] is geared to the verification of inconsistencies and maintenance of these ontologies. In its first uses, OBOL was intended to extract relations from OBO ontologies through linguistic analysis of names and definitions of Gene Ontology terms. In order to extract "is_a" relations, for instance, Mungall (2004) adopted three rule patterns based on the concepts of Genus-Differentia:

G, Q = (D) is_a G, Q = (D') if D is_a D'

G, Q = (D) is_a G', Q = (D) if G is_a G'

G, Q = (D) is_a G if ¬(∃D' such that D is_a D')

where G denotes *Genus;* Q = (D) denotes *differentia*, D denotes concepts in *differentia* (features), the comma means the logical AND.

These rules can be better understood with examples [Mungall 2004]:

(1) G,Q = (D) is_a G,Q = (D') if D **is_a** D'

Ex.: chromoplast membrane **is_a** plastid membrane, because

G ⟶ membrane,

Q = (D) ⟶ (*differentia* = chromoplast) ⟶ D = chromoplast,

Q = (D') ⟶ (*differentia* = plastid) ⟶ D' = plastid, and

chromoplast **is_a** plastid

We can say there is a subtype (is_a) relation between two different concepts of the same genus if there is a subtype relation between the features that distinguish the two concepts. In the case of the example, *chromoplast membrane* and *plastid membrane* are of the same genus (*membrane*), but they differ in the type of the *membrane*. In order for us to state that one of the terms is a subtype of the other, we need that one of the features that provide the difference among the concepts be in a subtype relation amongst the other. In the example, since both terms are of the same genus (*membrane*), we can state that *chromoplast membrane* is_a *plastid membrane* if a *chromoplast* is_a *plastid*.

(2) G,Q = (D) is_a G',Q = (D) if G **is_a** G'

Ex.: vitamin E biosynthesis **is_a** vitamin E metabolism, because

G ⟶ biosynthesis,

G' ⟶ metabolism,

Q = (D) ⟶ (*differentia* = vitamin E), and

biosynthesis **is_a** metabolism

Two concepts of different genus that contain the same differentia can have a subtype relation if one of these genus are subtype of the other. In the example, the differentia of the two terms regarding its generic term (their genus) is the same (*vitamin*). So, in order for one of the terms be a subtype of the other, we need both genus (*biosynthesis* and *metabolism*) to have a subtype relation. In this case, since both terms have the same differentia (*vitamin*), we can say that *vitamin E biosynthesis* is_a *vitamin E metabolism* if *biosynthesis* is_a *metabolism*.

(3) G,Q = (D) **is_a** G se ¬(∃D' such that D is_a D')

Ex.: primary septum **is_a** septum, because

G ⟶ septum, and

Q = (D) ⟶ (*differentia* = primary)

If there is no concept in the ontology that represents the differentia between the two concepts, then the term is directly subordinated to the term that represents the genus. In the example above, there is no term in the ontology that represents the differentia of *primary septum* (that would be *primary*). Thus, we can say that primary septum is directly related to the term that represents the genus of *primary septum*, which is *septum*.

More recently, OBOL was improved[4] and is being used as an OBO tool to accomplish the "cross product" between ontologies, aiming to find errors and implicit relations between OBO ontologies [Mungall et al. 2011]. However, OBOL's new version is concentrating a considerable part of its functionalities looking for potentially new relations (such as "occurs_in" and "adjacent_to"), in detriment of more basic relations (such as "part_of" and "is_a"). These new relations may enhance the ontologies by providing additional semantics to their terms, allowing the user to visualize their hierarchical (flat) structure through the relations is_a and part_of, and also to visualize their spatial structure through the relations adjacent_to and occurs_in.

[4] `http://wiki.geneontology.org/index.php/Obol/`

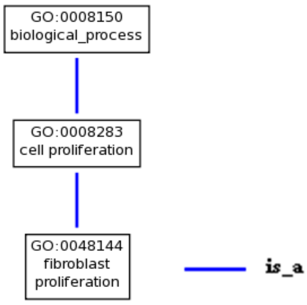


Fig. 2. Hierarchy of the term fibroblast proliferation (GO:0048144)

Our work is carried out according to several OBO initiatives, such as OBOL, which have the goal to improve ontologies quality by rescuing implicit information embedded in them. Initially, we have directed our efforts towards the is_a relation because, as explained before, this relation is the most important, once it provides a semantic foundation by relating concepts to sub-concepts in the actual ontology hierarchical structure [Brachman 1983]. Flaws and omissions of this kind of relation lead to several problems in researches that use them, and may compromise some of their results [Smith et al. 2005; Kohler et al. 2006]. Fig. 2, obtained from AmiGO[5] search interface, illustrates the representation of the hierarchy of fibroblast proliferation (GO:0048144), a Gene Ontology term. Fig. 2 shows that the is_a relation is the main relation responsible for the hierarchy of this term.

5. A SEMI-AUTOMATIC APPROACH TO EXTRACT NEW RELATIONS IN ONTOLOGIES

In the previous section, we presented methods and strategies for information extraction and also some related work. By building upon these works and analyzing biomedical ontologies, we developed an information extraction approach to be used on ontologies and other databases. The approach works with term nomenclature and definition, aiming to enrich ontologies. It uses multiple extraction strategies and several mechanisms to provide support to the process of information extraction, such as the linguistic tools GATE [Bontcheva et al. 2003] e NLTK [Bird et al. 2009]. This approach is implemented through the mechanism EI-ONTO, which provides support for all the steps of our approach.

In the context of the current work, our approach is applied to biomedical ontologies to extract hierarchical "is-a" relations. The main input to our approach is the term names and definitions. However, with some adaptations, the approach can be used for different purposes and on different data sources. As differentials for our approach we have:

—It has a "knowing the corpus" macro-step, prior to the information extraction step, allowing a probable precision gain in that step;

—It can be adapted to other domains and to find other types of relations;

—It uses multiple extraction strategies (dictionary-based, rule-based, machine learning-based, statistics-based) while other approaches use only one or two strategies.

—It has more interaction with users than other approaches. The user interacts with the approach in almost all steps and sub-steps. This allows errors to be found in the initial stages, avoiding rework.

Fig. 3 illustrates the proposed approach, which is divided into two macro-steps: knowing the corpus and working the corpus. Notice that the ontology is an input to the approach, and auxiliary bases

[5] http://amigo.geneontology.org/

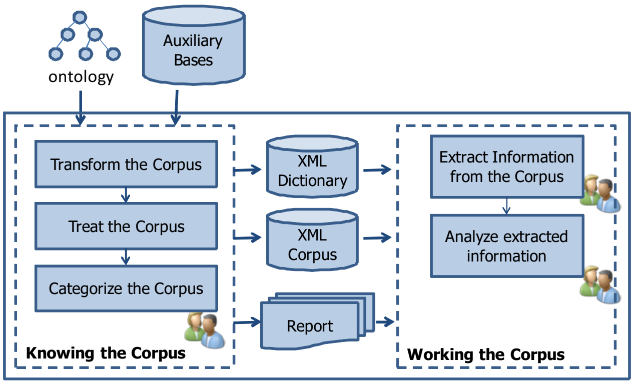


Fig. 3. Approach to Extracting Information in Ontologies

(such as a data dictionary) can also be used to help in the information extraction process. Also, some of the steps use input from domain specialists (categorize the corpus, extract information from the corpus and analyze extracted information).

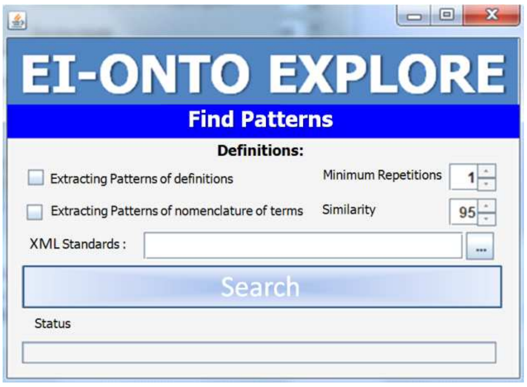
The first macro-step has the goal of studying the corpus and is divided into three steps: (i) transform the corpus; (ii) treat the corpus; and (iii) categorize the corpus.

Transform the corpus. This step has the objective of cleaning and transforming the corpus and the auxiliary bases in two XML documents (XML Corpus and XML Dictionary). The XML corpus contains a summary of the ontology that will be treated (concepts, their synonyms and definitions), and the XML Dictionary contains the auxiliary bases which will support the process of information extraction that uses the dictionary-based strategy. This dictionary is used on the second step, to help marking the occurrence of concepts inside definitions.

Treat the corpus. This step has the objective of applying natural language processing techniques (part-of-speech tagging, lemmatizing, tokenization of words and sentences, parsing, etc) with the intention to facilitate the information extraction process. For this, linguistic tools like GATE and NLTK are used. In our approach, these two mechanisms are used, because the tests we conducted to choose linguistic tools indicate that both tools have advantages and disadvantages. For example, the GATE tool is best suited for sentences tokenizing, while NLTK is best suited for part-of-speech tagging. Considering this, we have decided to use them together, hoping to extract the best of each one.

Categorize the corpus. This step is subdivided in two, and it is, in fact, where corpus recognition occurs. The first sub-step is to Calculate the Relevance of Corpus Elements. It uses the statistics-based strategy and aims at the recognition of delimiters concepts and verbs that are relevant to the domain. As default, we use absolute frequency as the relevance score of a given element. At the end of this sub-step two lists are generated: one containing all the verbs ordered by its frequency, and another one containing all the domain concepts delimiters, i.e., the words that occur more frequently after a concept, in a definition. The first list is used by the user of the approach to verify which verbs can be used to find the target relation in the domain. The second list is used to help us find new concepts in the corpus. This second list is necessary because the tools used in linguistic analysis presented difficulties to tag the biomedical corpus, and were unable to identify with precision the names of the concepts. By using the second list, we were able to obtain more accuracy on tagging the corpus concepts.

The second sub-step is to Find Corpus Patterns. It uses a machine learning strategy and aims at finding patterns in the definition and in the nomenclature of terms. Such patterns can be applied on information extraction. The algorithm used to find patterns in our approach is based on the technique



EI-ONTO EXPLORE

Find Patterns

Definitions:

□ Extracting Patterns of definitions — Minimum Repetitions: 1

□ Extracting Patterns of nomenclature of terms — Similarity: 95

XML Standards :

Status

Fig. 4. EI-ONTO Interface

of Edit-Distance [Cormode and Muthukrishnan 2007], i.e., it verifies the similarity between two sets of words, considering how many insertions or removals have to be made in one set of characters so that it becomes equal to the other. We use a sample set as the training base for this algorithm. After the training phase, the algorithm groups the definitions and names of the concepts in *clusters* by means of a similarity score defined in the EI-Onto mechanism, which implements the tool. At the end of this step, we generate a list containing the identified patterns. Each pattern describes a cluster that was found by the algorithm. The generation of the patterns is (also) made by an algorithm similar to the Edit Distance algorithm. The patterns are then validated by a specialist who will decide which ones should be used for information extraction. These processes and the identified patterns are better visualized in the next section.

Fig. 4 shows the interface built to identify the patterns. In this part of the system it is possible to choose the similarity threshold and also the minimum number of repetitions for a string to be considered a pattern. In other words, this parameter represents the minimal cardinality that a cluster must have in order to be considered a pattern. Moreover, it is also possible to search for predefined patterns inserted in XML format. As an example, we can use the one predefined by the Gene Ontology Consortium, as, for example:

[x] development

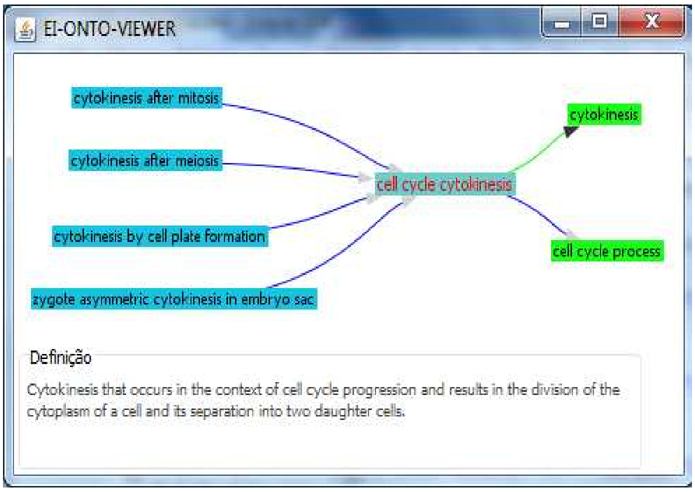
Standard definition: The process whose specific outcome is the progression of the [x] over time, from its formation to the mature structure.[6]

This pattern illustrates the standard way to describe all terms related to a development in Gene Ontology. Although these patterns have been initially developed for use in GO, they are also applied to several other ontologies in the OBO Consortium.

The second macro-step of our approach has the goal of extracting information from the corpus and uses the output data produced by the previous step. It is divided into two steps: (i) extract information from the corpus; and (ii) analyze extracted information.

Extract information from the corpus. This step has the objective of extracting information from the corpus and it is subdivided into three steps. The first sub-step aims at adapting the algorithms used on the approach to extract information according to the patterns found in the first macro-step. This sub-step is important to adapt the approach according to the specificities of each ontology or database that is being analyzed. The algorithms adapted in this step are divided in two kinds: those who deal with terms nomenclature, and hence were adapted from patterns that are related to terms nomenclature; and those related to the definition field, which were adapted from patterns related

[6] `http://www.geneontology.org/`


EI-ONTO-VIEWER
cytokinesis after mitosis
cytokinesis after meiosis
cytokinesis by cell plate formation
zygote asymmetric cytokinesis in embryo sac
cell cycle cytokinesis
cytokinesis
cell cycle process

Definição

Cytokinesis that occurs in the context of cell cycle progression and results in the division of the cytoplasm of a cell and its separation into two daughter cells.

Fig. 5. EI-ONTO VIEWER

to the definition field. The second sub-step aims at applying the algorithms with the purpose of extracting information, taking advantage of terms nomenclature. The third sub-step aims at applying the information extraction algorithms related to the terms definition. In this step we use the rules and dictionary-based strategies. A strategy based on rules is adopted for the identification of implicit information, according to the patterns found. A strategy based on dictionary is used to tag the corpus of concepts expressed in the XML dictionary.

Analyze extracted information. In this step, the extracted information is validated. The validation allows several analysis to be done. For instance, it allows the validation of the terms hierarchy, the verification of consonance of information expressed in the ontologies with what is implicit in the definition field and the in term nomenclature, etc. This validation occurs with the help of reporting and visualization mechanisms such as EI-ONTO VIEWER, which uses the information visualization toolkit Prefuse[7]. Fig. 5 shows part of an ontology in the mechanism EI-ONTO VIEWER. In this mechanism the new relationships (the ones detected by our approach) are represented in different colors allowing a better assessment by the specialist. In Fig. 5, the selected visualized term has its font color in red. Terms highlighted in blue with font color in black represent the direct descendant terms of the visualized term. The terms highlighted in green with font color in black are the direct ancestor terms of the visualized term The relations in blue (represented by blue arrows) are is_a relations already existent in the ontology and the relation in green (represented by a green arrow) is the is_a relation found by our approach. We have implemented EI-ONTO VIEWER as part of our approach.

6. EXPERIMENT AND RESULTS

We have applied our approach in an experiment involving two OBO ontologies: Gene Ontology (Biological Process branch) and INOH Event. The former was chosen for being the most used OBO ontology in genome annotation experiments nowadays and the later because it has several aspects in common with the Gene Ontology Biological Process branch. The experimental evaluation was divided in two phases. In the first phase, the approach was applied in the two ontologies. In this phase, after the validation by two biologists, a first result was obtained: there is useful implicit information in the ontologies and this information could be used to complement existing knowledge in the ontologies. In the second phase, two alignments were made: one with the two original ontologies,

[7] http://prefuse.org/

i.e., without any additional information (Fig. 6), and the other with the ontologies enriched by the information extracted by our approach (Fig. 7). In this second phase, we could highlight the quality of the information extracted by our approach, through the improvement we obtained in the alignment process. Due to space restrictions, we refrain from describing more details about our experiment. The interested reader can refer to [Carvalho 2011] for a detailed discussion.

In the first part of the experiment we applied our approach in the Biological Process branch of GO and the INOH Event ontologies. After the application of the first step we have identified four patterns. These patterns are used in the second step of the approach to help finding new relations and are described as follows:

(i) **Pattern:** Substring

Description: If a term is a sub part of another term, it might have a hierarchical relation with that term or with another exact synonym of that term.

Example: alkane transport (GO:0015895) is_a transport (GO:0006810)

(ii) **Pattern:** Regulation

Description: All terms that have positive or negative regulation of [x] should possibly have in their hierarchy a relation with the term regulation of [x] or some synonym of regulation of [x].

Example: positive regulation of viral reproduction (GO:0048524) is_a regulation of viral reproduction (GO:0050792)

(iii) **Pattern:** OF-IN Pattern

Description: All terms that have the prepositions "of" or "in" should possibly have in its hierarchy a relation with the term that precedes those prepositions.

Example: localization of cell (GO:0051674) is_a localization (GO:0051179)

(iv) **Pattern:** First Term

Description: All terms having a definition that begins with the articles "a", "an", "the", "any or "those" might have in their hierarchies a relation with the first term expressed on the definition after those articles.

Example: **Term:** actin filament reorganization involved in cell cycle (GO:0030037)

Definition: The cell cycle process whereby rearrangement of the spatial distribution of actin filaments and associated proteins occurs.

Relation Found: actin filament reorganization involved in cell cycle (GO:0030037) is_a cell cycle process (GO:0022402)

Besides these patterns, we have verified that OBOL rules to find "is_a" relations could also be taken into consideration. Therefore, in the algorithms used by our approach, we have also included as patterns OBOLs rules 1 and 2 [Mungall 2004] (OBOL rule 3 is already covered by the substring pattern).

(i) **Pattern:** OBOL rule 1

Description: G,Q = (D) is_a G,Q = (D') if D is_a D'

A term is an "is_a" of a second term when the *genus* of the two are the same and the *differentia* of the first is_a *differentia* of the second term.

Example: catabolism by host of symbiont xylan (GO:0052365) is_a metabolism by host of symbiont xylan (GO:0052420)

Table II. Number of relations extracted using our approach

Pattern	*INOH Event* Total	*Biological Process* Total
Substring	31	436
Regulation	0	4
OF Pattern	29	15
OBOL Rule 1	45	7
OBOL Rule 2	1	216
First Term	2	55

Table III. Number of relations validated using our approach and precision of the extracted information

Pattern	*INOH Event* validated/total (Precision)	*Biological Process* (Precision)
Substring	21/31 (68%)	311/436 (71%)
Regulation	0/0 (-)	4/4 (100%)
OF Pattern	29/29 (100%)	14/15 (93%)
OBOL Rule 1	45/45 (100%)	2/7 (29%)
OBOL Rule 2	0/1 (0%)	125/216 (58%)
First Term	2/2 (100%)	55/55 (100%)

(ii) **Pattern:** OBOL rule 2

Description: G,Q = (D) is_a G',Q = (D) if G is_a G'

A term is an "is_a" of a second term when the *differentia* of the two are the same and the *genus* of the first is_a *genus* of the second term.

Example: o-glycoside catabolic process (GO:0016142) is_a o-glycoside metabolic process (GO:0016140)

(iii) **Pattern:** OBOL rule 3

Description: G,Q = (D) is_a G if ¬(∃D' such that D is_a D')

A term is an "is_a" of its Genus if its differentia is a qualifier for the term, ie the differentia cannot be a concept that can be expressed in the ontology. This pattern can be considered a subpattern of the pattern substring. Example: primary cell wall biogenesis (GO:0009833) is_a cell wall biogenesis (GO:0042546)

These patterns were used in the second macro step (working the corpus) with the objective of extracting relations. Table II shows the number of relations extracted by each pattern.

Afterwards, the relations we found were validated with the help of two biologists. Table III contains the number of extracted relations and, among those, the ones considered valid by the biologists. In this Table it is possible to verify that the approach has a high degree of precision. In pattern First Term, for example, 100% of the relationships found by the approach were correct. The precision of patterns Substring, OBOL Rule 1 and OBOL Rule 2 did not achieve 100%, because those patterns resulted in many part_of relationships, instead of is_a. The patterns with the lower precision were those inherited from the OBOL approach (OBOL Rule 1 and OBOL Rule 2), which achieved a precision of 29% and 58%, respectively for the Biological Process branch of Gene Ontology. For the INOH Event, the 0% of precision reflects a non expressive case, since the universe, in this case, was only 1 relation.

This first evaluation has shown that our approach is indeed very precise in correctly capturing new relations. However, we still needed to evaluate how useful this new extracted information is in practice. To do so, in the second phase we have set up an experiment where we measured how the new relations help in the alignment process of two ontologies, and consequently, in the reuse of those ontologies. The rationale in this case is that if the extracted information is useful, then the alignment will present improvements.

We have aligned these biomedical ontologies using the ontology alignment tool FOAM [Ehrig and

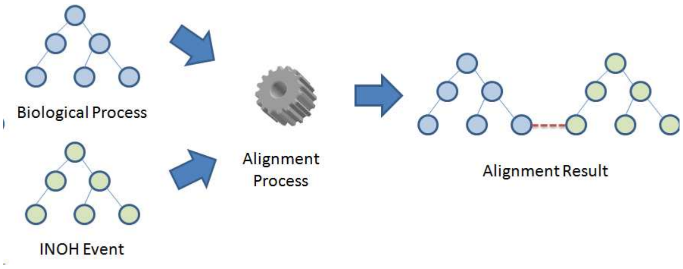


Fig. 6. First Alignment

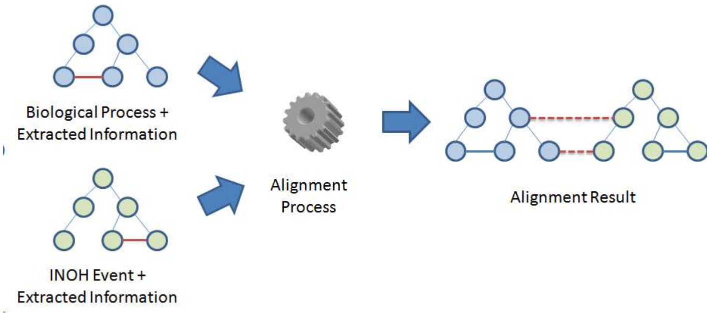

Biological Process + Extracted Information
INOH Event + Extracted Information
Alignment Process
Alignment Result

Fig. 7. Second Alignment

Sure 2005]. As described previously, we have made two different alignments, one with the original ontologies (Fig. 6) and the other with the enriched ontologies (Fig. 7). We call "enriched ontology" an ontology that has passed through our relation extraction process, and, as a consequence, has had new relations added to it.

The alignments were made using FOAM tool because in her work, Silva (2010) has made a thorough study about ontology alignment tools, choosing FOAM as the most viable to the alignment of biomedical ontologies. The parameters used in the alignment were:

—Type of alignment: Fully automated;
—Number of iterations: 10;
—Cut value: 0,97;
—Type of strategy: Decision Tree.

To classify the results we used the same methodology of Silva (2010). In her work, Silva (2010) also proposed a classification for the alignment results, according to a scale of 1 to 5 (Table IV), allowing intermediate scores for alignments (meaning partially correct alignments), instead of only considering them correct or incorrect. Although partially correct alignments cannot be considered as equivalence relations, they can mean other kinds of relations, such as "is_a", "occurs_in" and "part_of" [Silva 2010].

At the end of the experiment, results were revised by two biologists with experience in genome sequencing. The biologists have classified the aligned terms according to the classification presented in Table IV. In our results, every time there was a divergence in the values provided by the two biologists for a given term, we used the lowest value. For example, suppose the first biologist assigned degree 5 to a given pair of aligned terms, and the second biologist assigned degree 4 to the same pair of terms. In this case, we consider degree 4 for the aligned pair of terms.

After results validation, we have confronted each of the generated alignments. Results show a clear improvement on most results of the second alignment (the one enriched with the new identified relations), as shown in Tables V and VI and in Fig. 8.

Table IV. Classification for the alignment results (Source: [Silva 2010])

Degree	Explanation
5 (Correct)	The aligned pair of terms are equivalent
4 (Strong Relation)	The aligned pair of terms have a strong hierarchical relationship (is_a)
3 (Medium Relation)	The aligned pair of terms have a medium hierarchical relationship (is_a) or part_of relation
2 (Weak Relation)	The aligned pair of terms have a weak hierarchical relationship (is_a) or other type of relationship
1 (Incorrect)	The aligned pair of terms are unrelated

Table V. First Alignment validation results

Classification Degree	Amount	Percentage (%)
5 (Correct)	39	68%
4 (Strong Relation)	4	7%
3 (Medium Relation)	7	12%
2 (Weak Relation)	6	11%
1 (Incorrect)	1	2%

Table VI. Second Alignment validation results

Classification Degree	Amount	Percentage (%)
5 (Correct)	47	78%
4 (Strong Relation)	3	5%
3 (Medium Relation)	7	12%
2 (Weak Relation)	2	3%
1 (Incorrect)	1	2%

In Tables V and VI it is possible to detect an increase of correctly aligned pairs of terms (i.e., equivalent terms) in the second alignment. These tables also show a decrease in the number of aligned pairs of terms that were classified as having "weak relation" in the second alignment. It is also true that these tables show a decrease in the "strong relation". However, this occurred because one of the pairs of terms that was previously classified as having a "strong relation" in the first alignment could be correctly aligned in the second alignment. This is due to the new relations we added to the ontology, that FOAM was able to use in the alignment computation.

Our results (visualized graphically in Fig. 8) show a qualitative and quantitative increase in the number of correctly aligned pairs found. This result demonstrates that by complementing the ontologies with new relationships, it is possible to improve their reuse process. By improving the reuse process it is possible to offer more options to the annotator of terms, allowing a better choice of the term that will be used to annotate.

7. FINAL REMARKS

The motivation for the work presented in this paper was grounded on literature analysis and also on the evaluation of biomedical ontologies, where we found evidences of omission of relevant terms and relations, which could hinder the process of ontology reuse in the context of genome annotation. With the goal of trying to enrich information contained in ontologies, we have proposed an approach to enrich an ontology by extracting previously implicit knowledge contained inside the ontology, as for instance, in its terms nomenclature and definition. This approach is described with more details in Carvalho (2011).

Our approach was applied successfully on biomedical ontologies, and is grounded on several strategies, techniques and approaches on information extraction. In the scenario presented in this paper, the approach was applied to verify the existence of implicit information on biomedical ontologies,

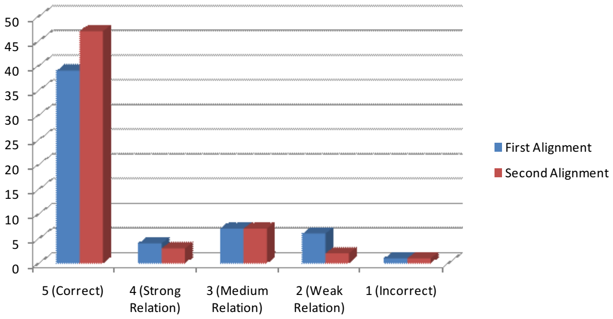


Fig. 8. Comparison of pair alignments in the first and second alignments

aiming to answer the following question: "Can implicit information contained in biomedical ontologies improve their reuse process?".

After an experimental evaluation, we have reached two major results. The first is an increase in the quality and quantity of the extracted information, showing that in fact there is implicit knowledge in biomedical ontologies, and that this knowledge can be made explicit. The second result relates to the usefulness of such knowledge, which was verified by the improvement on the quality of the alignment of two biomedical ontologies (GO Biological Process and INOH Event), after the extracted knowledge (new terms and relations) was used to enrich the ontologies.

Although much work is still needed to enhance the quality of biomedical ontologies, our proposal can be seen as a contribution on pointing ways to retrieve already existing, although implicit, knowledge. It is worth noting that our work is coming towards the efforts of OBO, which aims to constantly concentrate efforts on obtaining improvements and standardization in biomedical ontologies. As future work, we plan to adapt our approach to extract new terms, since many ontology descriptions refer to terms that are not explicit in the ontology. In addition, we will use the first macro step to recognize and extract new patterns of relationships. This will follow the line developed by Mungall et al., (2011).

REFERENCES

Agarwal, P. and Searls, D. B. Literature mining in support of drug discovery. *Briefings in Bioinformatics* 9 (6): 479–492, 2008.

Ananiadou, S. and McNaught, J. *Text mining for biology and biomedicine.* Artech House, 2006.

Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python.* O'Reilly Media, 2009.

Bontcheva, K., Maynard, D., Tablan, V., and Cunningham, H. Gate: A unicode-based infrastructure supporting multilingual information extraction. In *Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL).* Borovets, Bulgaria, pp. 1–8, 2003.

Brachman, R. What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer* 16 (10): 30–36, 1983.

Campos, M. L. A. *Em busca de Principios Comuns na Área de Representação da Informação: uma Comparação entre o Método de Classificação Facetada, o Método de Tesauro Baseado em Conceito e a Teoria Geral da Terminologia.* M.S. thesis, IBICT/Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 1994.

Carvalho, M. G. P. *Enriquecimento de Ontologias: Uma abordagem para Extração das Informações Implícitas.* M.S. thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2011.

Cormode, G. and Muthukrishnan, S. The string edit distance matching problem with moves. *ACM Transactions on Algorithms* 3 (1): 1–19, 2007.

Dahlberg, I. A referent-oriented, analytical concept theory for interconcept. *International Classification* 5 (3): 142–150, 1978.

Dahlberg, I. Conceptual definitions for interconcept. *International Classification* 8 (1): 16–22, 1981.

Ehrig, M. and Sure, Y. FOAM - Framework for Ontology Alignment and Mapping - Results of the Ontology Alignment Evaluation Initiative. In *Workshop on Integrating Ontologies.* Banff, Canada, pp. 72–76, 2005.

Fundel, K., Kuffner, R., and Zimmer, R. RelEx: Relation extraction using dependency parse trees. *Bioinformatics* 23 (3): 365–371, 2007.

Gomes, H. and Campos, M. L. A. Contribuição da teoria da terminologia na elaboração de hiperdocumentos com função de tesauro. In *Simpósio Iberoamericano de Terminologia (RITerm)*. Buenos Aires, Argentina, 2004.

Gu, H. H., Wei, D., Mejino, Jr., J. L. V., and Elhanan, G. Relationship auditing of the fma ontology. *Journal of Biomedical Informatics* 42 (3): 550–557, 2009.

Hakenberg, J., Leaman, R., Vo, N. H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C., and Gonzalez, G. Efficient extraction of protein-protein interactions from full-text articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7 (3): 481–494, 2010.

Kohler, J., Munn, K., Ruegg, A., Skusa, A., and Smith, B. Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics* 7 (212): 1–12, 2006.

Krallinger, M., Valencia, A., and Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology* 9 (2): S8.1–S8.14, 2008.

Mathiak, B. and Eckstein, S. Five steps to text mining in biomedical literature. In *European Workshop on Data Mining and Text Mining for Bioinformatics*. Pisa, Italy, pp. 47–50, 2004.

Michael, J., Mejino, J. L. V., and Rosse, C. The role of definitions in biomedical concept representation. In *American Medical Informatics Association (AMIA) Symposium*. Washington, DC, pp. 463–467, 2001.

Mungall, C. J. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics* 5 (6-7): 509–520, 2004.

Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J. I., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. Cross-product extensions of the gene ontology. *Journal of Biomedical Informatics* 44 (1): 80–86, 2011.

Ogren, P. V., Cohen, K. B., Acquaah-Mensah, G. K., Eberlein, J., and Hunter, L. The compositional structure of gene ontology terms. In *Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 214–225, 2004.

Ogren, P. V., Cohen, K. B., and Hunter, L. Implications of compositionality in the gene ontology for its curation and usage. In *Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 174–185, 2005.

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R., and Rhodes, S. Identification of Biological Relationships from Text Documents. In H. Chen, S. Fuller, C. Friedman, and W. Hershp (Eds.), *Medical Informatics Knowledge Management and Data Mining in Biomedicine*. Springer-Verlag, New York, USA, pp. 449–489, 2005.

Rinaldi, F., Schneider, G., Kaljurand, K., Dowdall, J., Andronis, C., Persidis, A., and Konstanti, O. Mining relations in the GENIA corpus. In *European Workshop on Data Mining and Text Mining for Bioinformatics*. Pisa, Italy, pp. 61–68, 2004.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., Konstanti, O., and Persidis, A. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine* 39 (2): 127–136, 2007.

Sales, L. F., Campos, M. L. A., and Gomes, H. E. Ontologias de domínio: Um estudo das relações conceituais. *Perspectivas em Ciência da Informação* 13 (2): 62–76, 2008.

Schulz, S., Kumar, A., and Bittner, T. Biomedical ontologies: what part-of is and isn't. *Journal of Biomedical Informatics* 39 (3): 350–361, 2006.

Silva, V. *Uma Abordagem para Alinhamento de Ontologias Biomédicas para Apoiar a Anotação Genômica*. M.S. thesis, PPGI, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2010.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consortium, T. O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25 (11): 1251–1255, 2007.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. Relations in biomedical ontologies. *Genome Biology* 6 (5): 1–15, 2005.

Smith, B., Köhler, J., and Kumar, A. On the application of formal principles to life science data: a case study in the gene ontology. In *Database Integration in the Life Sciences (DILS)*, E. Rahm (Ed.). Lecture Notes in Computer Science, vol. 2994. pp. 79–94, 2004.

Smith, B. and Kumar, A. Controlled vocabularies in bioinformatics: a case study in the gene ontology. *Drug Discovery Today: BIOSILICO* 2 (6): 246–252, 2004.

Smith, B. and Rosse, C. The role of foundational relations in the alignment of biomedical ontologies. In *World Congress on Medical Informatics (MEDINFO)*. San Franscisco, California, pp. 444–448, 2004.

Smith, B., Williams, J., and Schulze-kremer, S. The ontology of the gene ontology. In *American Medical Informatics Association (AMIA) Symposium*. Vancouver, British Columbia, pp. 609–613, 2003.

Wroe, C. J., Stevens, R., Goble, C. A., Ashburner, M., Wroe, C. J., Stevens, R., Goble, C. A., and Ashburner, M. A methodology to migrate the gene ontology to a description logic environment using daml+oil. In *Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 624–635, 2003.