Analysis and Classification of Bio-ontologies by the Structure of their Labels

Manuel Quesada-Martínez[1], Jesualdo Tomás Fernández-Breis[1], and Robert Stevens[2]

[1] Departamento de Informática y Sistemas
Facultad de Informática, Universidad de Murcia
CP 30100, Murcia, Spain
manuel.quesada@um.es, jfernand@um.es
[2] School of Computer Science, University of Manchester, UK,
robert.stevens@manchester.ac.uk

Abstract. The development and success of the Gene Ontology were key factors for attracting the interest of biomedical researchers and bioinformaticists to ontologies. In recent years, hundreds of biomedical ontologies have been produced, most of them developed in collaborative efforts and following a set of construction principles, including the use of a systematic naming convention and using descriptive labels. Such ontologies have been mainly used for supporting the annotation process, but more sophisticated uses would require such ontologies to have more axioms. In recent works, we have found that exploiting the structure of the labels could contribute to that axiomatic enrichment. Hence, in this work we perform a study of the labels of the ontologies available in BioPortal to classify them in terms of potential interest for their axiomatic enrichment.

Keywords: Biomedical ontologies; OWL; Ontology Engineering; Bioinformatics

1 Introduction

The development and success of the Gene Ontology (GO) [1] were key factors for attracting the interest of biomedical researchers and bioinformaticists to ontologies. Many projects have used GO for supporting the annotation of biomedical data and as an instrument for functional analysis. As a consequence, many bio-ontologies have been developed, trying to emulate the impact of GO, but in different biomedical subdomains. Most of such ontologies have been developed using the design principles of the OBO Foundry and are available at BioPortal [4].

Many of these ontologies have not been created by ontology engineers, but by domain experts. This should help the veracity of the domain knowledge, but not necessarily the engineering of the ontology. Many such ontologies are plain taxonomies and controlled vocabularies, so they have a lower degree of axiomatization. Many ontologies do, however, have much information within the labels

and textual definitions of the classes [9]. These are useful for human users, but not much good for machine processing. The transformation of axiomatically lean ontologies into axiomatically rich ones is an interesting task [3]. This particular axiomatic enrichment process was based on the processing of the structure of the labels of biomedical ontologies. In addition, the results presented in [6] showed that the structure and content of the labels of the classes of relevant biomedical ontologies like GO or SNOMED-CT [8] were suitable for application of the enrichment process proposed in [3] and for which some degree of automation is being provided by a software tool that is currently being developed in our research group [5]. Thus, exploiting the semantics within the labels on classes of such ontologies would offer significant benefits, because axiomatically rich ontologies would be helpful for advanced semantic analysis of biomedical data. The analysis methods could not only exploit the labels and the taxonomic links, but also a series of properties that would provide new analysis dimensions.

Hence, in this paper we will perform a systematic analysis of the ontologies publicly available in BioPortal [4] with the objective of identifying which ones are more suitable for the application of the enrichment approach using the OntoEnrich. Thus, our objective is to analyse and classify the BioPortal ontologies by the properties of the structure of their labels.

2 Lexical patterns

2.1 The concept and structure of lexical patterns

A lexical pattern is an ordered group of words (tokens) that is repeated in the labels of classes in an ontology. Lexical patterns are characterized in our approach by its content, length (number of tokens), frequency, whether the lexical pattern corresponds to the full label of a class, and whether some of its tokens are found in other ontologies either as classes or properties. Our approach assumes that groups of words that appear in many labels are likely to encode some domain meaning. For instance, the lexical pattern *binding* appears in 4.38% of the labels of Gene Ontology classes. Examples of such labels are *vitamin binding* and *isoprenoid binding*, which stand for the binding of two particular types of chemical substances, namely, vitamins and isoprenoids. Both labels have similar structure, but the ontology does not contain an axiomatic description of what binding, vitamin or isoprenoid mean.

The approach uses two additional notions, namely, sub-pattern and super-pattern. First, a lexical sub-pattern is a sub-sequence of a pattern. Second, a lexical super-pattern of a given pattern is a pattern that includes the latter one. From a given pattern, super-patterns can be obtained by extending the pattern in any direction. In our running example, the lexical pattern *binding* is a sub-pattern of *vitamin binding*. Similarly, the lexical pattern *activity* (23.28%) can be extended to the super-pattern *oxidoreductase activity, acting on peroxide as acceptor*, which captures a considerable amount of knowledge, but not in a way machines could take advantage of it.

Nevertheless, not every group of repeated words embed domain knowledge. Text based approaches usually make use of lists of stop-words to identify words without meaning and, therefore, useless for further processing. In our approach, we filter out all the patterns that consist of only stop-words. This is not the only filtering mechanism included in our approach, since it can also remove all the patterns whose frequency is not high enough. For this purpose we use the *coverage of a lexical pattern* that is the minimum percentage of classes where a lexical pattern must appear to be included in the result.

If we analyse the structure of the labels of the families of functions in GO, we could identify some regularities. Most binding functions have a label with the structure X *binding*, where X is a chemical substance. Most types of structural molecule activities have a label with the structure *structural constituent of* Y, where Y is a macromolecular complex. Given such regularity, we should be able to systematically pull out patterns of axioms that can make the semantics explicit. Such naming conventions might be exploited to generate patterns of axioms that make the information in the label computationally explicit.

2.2 Organization of labels and detection of lexical patterns

The time required in the detection and navigation through the hierarchy of lexical patterns increases as the number of labels of the ontology grows. For this reason, the use of an accurate organization of the tokens in a label is required in order to tackle ontologies with thousands of labels and words.

Our method represents the map of patterns as a graph of tokens, which is built as the input ontology is processed. The exploitation of such a graph permits the identification of the lexical patterns and links between them. Each node of the graph corresponds to one token (our approach currently uses blanks as delimiters). In addition, each node may be linked to other nodes through an arrow. Such arrows mean that such nodes appear consecutively and in that order in a label. Apart from these elements, the graph also includes information about other issues such as the position in the label (e.g., the same word could appear several times in the same label), the URI of the class, and so on. All these features have motivated the use of our own representation of the graph. Figure 1 shows a graph where four labels of last example are represented.

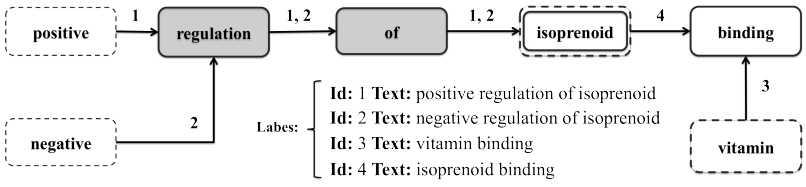

Fig. 1. Brief representation in the graph of labels with the content of GO example

2.3 Detecting concepts and properties from other ontologies

The re-use of content from already existing ontologies has traditionally been considered good practice [7]. Enrichment processes should also benefit from the fact that there might be words of the labels that are names/labels of classes and properties in other bio-ontologies. Our method is capable of looking for matches in external ontologies for the content of the lexical patterns. Our hypothesis is that such matches would reinforce the interest of a particular lexical pattern and the idea that they might be embedding domain knowledge. For instance, if *vitamin* is a class of the ChEBI ontology [2] (`http://purl.obolibrary.org/obo/CHEBI_33229`), having the possibility of re-using such class when enriching the Gene Ontology class whose label is *vitamin binding* would be useful for the ontology builder. Given that the content of such an external ontology will certainly be created by domain experts, we would be enriching the axiomatic content of the ontology where the lexical patterns are analyzed enhancing its quality.

For this purpose we define two types of matches, namely, exact and partial. An exact match of a given lexical pattern happens when the class of an external ontology has the same label. A partial match happens when the lexical pattern is contained in the label of an external class or when the opposite situation stands. Finally, it should be noted that matches must be considered as suggestions, since no logical equivalence is computed.

3 Design of the Analysis of the BioPortal repository

BioPortal [4] is the largest repository of biomedical ontologies. Users can submit their ontologies for publication in BioPortal; during the submission process the creator of the ontology specify meta-information such as the ontology name, description, abbreviation, format, version and so on. Concerning the management of versions in the repository, once users have developed a new version of an ontology, it is added to the repository being available all of the previous uploaded versions. At the time of writing, BioPortal contains more than two hundred biomedical ontologies and controlled vocabularies. We wonder if BioPortal ontologies follow the design criteria proposed by the OBO Foundry, which include to use a systematic naming convention and to have labels understandable by humans. On the contrary, the axiomatic richness of their ontologies is limited. Thus, analysing the content and structure of their labels with the goal of identifying which ones are more suitable for applying enrichment processes is relevant.

3.1 Description of the experiment

The experiment consists of executing our method on each BioPortal ontology. For each ontology, its lexical patterns will be extracted and analyzed. Then, the results will be analyzed in order to classify the ontologies in groups of interest.

Our objective is to assign to each group a degree of interest for having ontology enrichment methods applied. This grouping will be achieved with clustering techniques and the main variables that will be used in the analysis are:

- *Number of labels*: number of labels in an ontology.
- *Number of lexical patterns*: number of lexical patterns found in an ontology.
- *Coverage of lexical patterns*: frequency threshold for a lexical pattern to be considered. The analysis of the labels will be done with different values.
- *Classes affected by lexical patterns*: this issue must be considered in both absolute and relative terms, so they would be two different variables in terms of the analysis. Both variables would stand for the number and percentage of classes in which lexical patterns are found.
- *Classes affected by matches*: this issue must be considered in both absolute and relative terms, so they would be two different variables in terms of the analysis. Both variables would stand for the number and percentage of classes for which exact matches are found.
- *Repetition of words*: for each ontology, we will obtain how many different words exist in the labels.

4 Results

We worked with the BioPortal ontologies publicly accessible in OWL format in November 2012. Our corpus consisted of 178 OWL ontologies, from which 19 were discarded due to their importing inaccessible OWL files; another 41 were discarded due to the absence of labels on their entities. We have analyzed the labels of the ontologies with different values for the coverage threshold: from 0.2% to 1.0% with increments of 0.2 and from 1.0% to 5.0% with increments of 1.0. In this paper we show the results with the coverage set to 1%, but the complete results are available at `http:\\miuras.inf.um.es/biotest`.

4.1 Description of the ontologies by their labels

Table 1 shows the global descriptors for the data set composed of 118 ontologies. The first columns focus on metrics about the lexical patterns. First, we show the total number of lexical patterns and how many are unique. In this way, we find that 20.19% of the patterns appear in more than one ontology so they could be re-used in multiple enrichment processes. Next, we show the mean and maximum values for the length of patterns and frequency, and the percentage of lexical patterns for which matches in the same or external ontologies are found. The fact of finding more matches in external ontologies suggests that the re-use of this content for enriching current biomedical ontologies may be a significant contribution. Furthermore, we have calculated that the mean value of repeated words in labels is 67.7%, ranging from 50.01% to 94.7% and this can be interpreted as a sign of regularity.

Figure 2 shows the percentage of lexical patterns (out of the total number of labels in an ontology) and the percentage of classes that they cover for each

Table 1. Numerical metrics about the features of the lexical patterns

Number		Length			Frequency			Class Matches	
Total	Unique	Max	Mean	X50	Max	Mean	X50	Source	External
8 175	6 254	12	2.01	1	47 623	83.12	4	15.60%	36.44%

ontology analyzed. In general, both variables have a similar trend, except for those cases where the lexical patterns exceed the number of labels. Statistically speaking, these variables show a positive correlation (r=0.938, p=0.000).

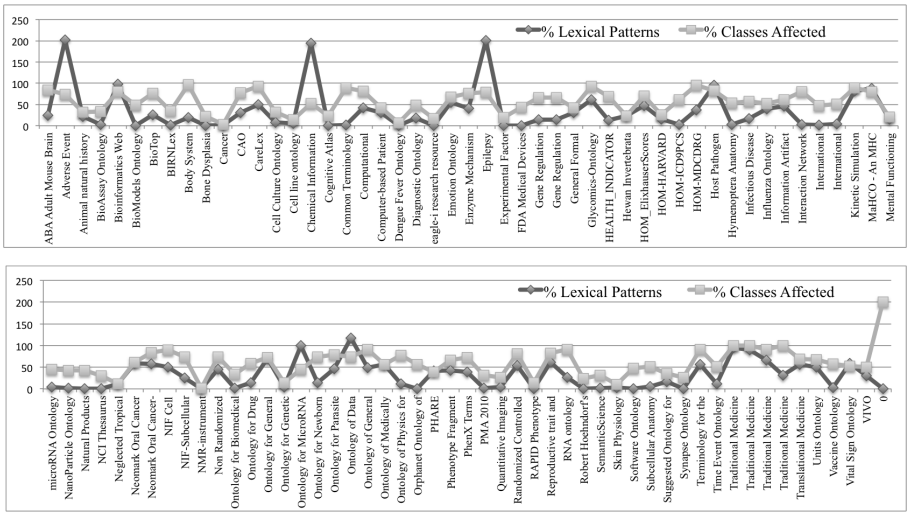

Fig. 2. Relation between the number of lexical patterns and the number of classes for whose labels patterns are found

Figure 3 shows the number of lexical patterns with one or more exact matches and their relation with the number of lexical patterns for each ontology. As expected, there is also a positive correlation (r=0.765, p=0,000). Finally, a positive correlation between the number of lexical patterns and the repetition of words is significant (r=0.41, p=0,000). However, the correlation between the number of classes affected by the patterns and the repetition of words is found negative but not significative (r=-0,118, p=0,246).

4.2 Classifying the ontologies

We have grouped the set of ontologies in clusters. We performed first an agglomerative hierarchical cluster. The inspection of the dendrogram suggested

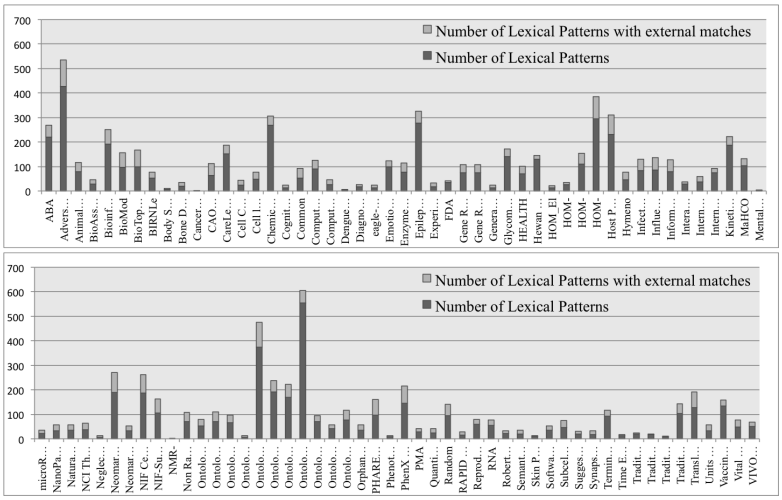

Fig. 3. Number of lexical patterns with exact matches compared with the total number of lexical patterns

the existence of three differentiated groups of ontologies. We then selected the clustering variables: percentage of classes for which patterns have been found, percentage of classes for which matches have been found, and percentage of repetition of words. We have applied the k-means clustering for such variables, obtaining three groups, whose centroids have the values shown in Table 2.

Table 2. Description of the clusters

Variable	**Cluster1**	**Cluster2**	**Cluster3**
Patterns	81.73	20.85	50.03
Matches	19.16	2.63	9.90
Repetitions	72.96	62.85	73.45
Ontologies	42	24	33

Cluster1 includes ontologies whose percentage of classes with patterns is high and those with a highest percentage of classed affected by matches. The ontologies of this cluster are the most suitable for applying enrichment methods. Cluster2 includes ontologies with low scores for both patterns and matches and, therefore, it includes the least suitable ones, that is, the ones for which the enrichment process would be less effective. Finally, Cluster3 includes ontologies with an intermediate score for class labels formed from patterns. The 99 ontologies for which patterns were found have been distributed as shown in Table 2: 42 in Cluster1, 24 in Cluster2 and 33 in Cluster3. Hence, it could be said that it seems interesting to apply enrichment processes to the members of Cluster1 and Cluster3. The members of each cluster are listed at `http:\\miuras.inf.um.es/biotest`.

5 Conclusions

In this paper we have systematically analyzed the OWL ontologies publicly available in BioPortal. The objective of the analysis was to study the structure of the labels to detect which ontologies should be worth axiomatically enriching by applying semi-automatic processes. A series of variables have been measured for every ontology and the values obtained have been analyzed in different ways, providing information of interest about the ontologies. The results of the clustering method allowed us to classify the ontologies in groups of interest. These results should be taken into account when deciding whether to enrich an ontology by exploiting the structure of its labels.

Acknowledgments. This project has been possible thanks to the funding of the Spanish Ministry of Science and Innovation through grant TIN2010-21388-C02-02 and co-funded by FEDER. Manuel Quesada-Martínez is funded by the Spanish Ministry of Science and Innovation through fellowship BES-2011-046192.

References

1. G. O. Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 23(May):25–29, 2000.
2. P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. Chemical entities of biological interest: an update. *Nucleic acids research*, 38(Suppl 1):D249–D254, 2010.
3. J. T. Fernandez-Breis, L. Iannone, I. Palmisano, A. L. Rector, and R. Stevens. Enriching the gene ontology via the dissection of labels using the ontology pre-processor language. In *Proceedings of the 17th international conference on Knowledge engineering and management by the masses*, EKAW'10, pages 59–73, Berlin, Heidelberg, 2010. Springer-Verlag.
4. N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. D. Storey, C. G. Chute, and M. A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web-Server-Issue):170–173, 2009.
5. M. Quesada-Martínez, J. T. Fernández-Breis, and R. Stevens. Enrichment of owl ontologies: a method for defining axioms from labels. In *Proceedings of the First International Workshop on Capturing and Refining Knowledge in the Medical Domain (K-MED 2012)*, pages 1–10, 2012.
6. M. Quesada-Martínez, J. T. Fernández-Breis, and R. Stevens. Extraction and analysis of the structure of labels in biomedical ontologies. In *Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems*, MIXHS '12, pages 7–16, New York, NY, USA, 2012. ACM.
7. A. L. Rector, S. Brandt, N. Drummond, M. Horridge, C. Puleston, and R. Stevens. Engineering use cases for modular development of ontologies in owl. *Applied Ontology*, 7(2):113–132, 2012.
8. M. Stearns, C. Price, K. Spackman, and A. Wang. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.
9. A. Third. "Hidden semantics": what can we learn from the names in an ontology? In *7th International conference on Natural Language Generation*, 2012.