Lexical Characterization and Analysis of the BioPortal Ontologies

Manuel Quesada-Martínez, Jesualdo Tomás Fernández-Breis
Departamento de Informática y Sistemas Facultad de Informática, Universidad de Murcia CP 30100, Murcia, Spain. manuel.quesada@um.es, jfernand@um.es

Robert Stevens
School of Computer Science, University of Manchester, UK.
robert.stevens@manchester.ac.uk

Abstract. The increasing interest of the biomedical community in ontologies can be exemplified by the availability of hundreds of biomedical ontologies and controlled vocabularies, and by the international recommendations and efforts that suggest ontologies should play a critical role in the achievement of semantic interoperability in healthcare. However, many of the available biomedical ontologies are rich in human understandable labels, but are less rich in machine processable axioms, so their effectiveness for supporting advanced data analysis processes is limited. In this context, developing methods for analysing the labels and deriving axioms from them would contribute to make biomedical ontologies more useful. In fact, our recent work revealed that exploiting the regularities and structure of the labels could contribute to that axiomatic enrichment.

In this paper, we present an approach for analysing and characterizing biomedical ontologies from a lexical perspective, that is, by analysing the structure and content of the labels. This study has several goals: (1) characterization of the ontologies by the patterns found in their labels; (2) identifying which ones would be more appropriate for applying enrichment processes based on the labels; (3) inspecting how ontology re-use is being addressed for patterns found in more than one ontology. Our analysis method has been applied to BioPortal, which is likely to be the most popular repository of biomedical ontologies, containing more than two hundred resources. We have found that there is a high redundancy in the labels of the ontologies; it would be interesting to exploit the content and structure of the labels of many of them and that it seems that re-use is not always performed as it should be.

Keywords: Biomedical ontologies, OWL, Ontology Engineering, Bioinformatics.

1 Introduction

Many biomedical ontologies have now been developed, stimulated by the increasing importance of biomedical ontologies in the scientific community. In fact,

important challenges like semantic interoperability in healthcare consider ontologies fundamental as stated in the SemanticHealth final report [2] and SemanticHealthNet (http://www.semantichealthnet.eu). 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.

BioPortal [4] is likely to be the most important repository of biomedical ontologies, and contains more than three hundred biomedical ontologies and controlled vocabularies so far and such knowledge resources come from a variety of ontology builders. Consequently, the analysis and characterization of the proper-ties of BioPortal ontologies becomes relevant in order to allow users and developers of biomedical ontologies to know what they can expect from such ontologies. Many Bioportal ontologies are related to the OBO Foundry (http://www.obofoundry.org). The OBO Foundry has developed a series of criteria that developers of biomedical ontologies should use to contribute to the development of an orthogonal collection of biomedical ontologies (http://obofoundry.org/crit.shtml). Such a collection of ontologies should benefit from the re-use of the content produced in already existing ontologies and be human and machine friendly. On the human's side, the OBO Foundry proposes to use a systematic naming convention, thus ontologies should have meaningful labels and be well documented. However, as shown in [7], this is not a specific property of OBO ontologies, but found in many available ontologies. On the machine side, ontologies should be machine processable and labels are not very useful for this. The benefit we can expect from the machine processing of the ontologies depends on the axiomatic richness of the ontologies and on other factors related to the quality of the ontology, discussion which is out of the scope of the present paper. However, according to our experience with biomedical ontologies, such richness is limited. Many such ontologies are no more than plain taxonomies and controlled vocabularies, so they have a lower degree of axiomatisation.

Thus, methods for the axiomatic enrichment of biomedical ontologies would permit ontologies to be computationally more powerful. Given that biomedical ontologies are supposed to be rich in labels, our working hypothesis is that the content and structure of such labels can be useful information for supporting the axiomatic enrichment. By structure of the labels, we mean the regularities that can be found in the groups of words that form such labels.

The richness of the labels means that the corresponding texts may be encoding biomedical domain knowledge. Consequently, their study should be useful for de-riving domain knowledge and enriching the axiomatic definition of the ontology classes.

For example, the expression *negative regulation* stands for the prevention or reduction of, generally, a biological process. This linguistic expression appears in several biomedical ontologies. On the one hand, the lack of axioms would only permit machine to exploit the labels but not the biological meaning of the concept. On the other hand, it is not guaranteed that all the ontologies in which *negative regulation* is found share the axioms, if any, for this concept.

In previous work [6], we showed that the labels of the classes of relevant biomedical ontologies like GO [1] or SNOMED-CT (http://www.ihtsdo.org/snomed-ct/) were suitable for application of the enrichment process proposed in [3]. In the current paper, a systematic analysis of the labels of BioPortal ontologies is performed, whose objectives are: (1) characterizing the ontologies by their labels; (2) identifying which ontologies are more suitable for applying enrichment processes; (3) analyzing whether ontology content is re-used in an appropriate way in existing biomedical ontologies. Such a study will provide new insights about biomedical ontologies and will drive our research on the enrichment of such ontologies.

2 Methods

2.1 Representation and Extraction of Lexical Patterns

Our basic assumption is that groups of words that appear in many labels are likely to encode some domain meaning. We call such groups of words (or tokens) lexical patterns. In our approach, a lexical pattern has some basic descriptors associated, like its content, length (number of tokens), or frequency in an ontology. We are going to illustrate these concepts using the *Vaccine Ontology* (http://bioportal.bioontology.org/ontologies/49452). Its lexical pattern of length one *virus* appears in 25.57% of the labels. On the other hand, the lexical pattern *preparation of* has length 2 and it appears in 14.34% of the labels. Other examples are *canine*, *Rhinotracheitis-Virus*, *protein vaccine*, *Modified Live virus* and so on.

When analyzing the labels of a particular ontology, we are also interested in knowing which lexical patterns correspond to the full label of a class, and which ones are contained in the labels of external ontologies. If a lexical pattern corresponds to a full label of a class, it might be encoding the meaning of a domain concept and there should be a relation between this class and the other labels in which the pattern is found. Moreover, the axioms extracted from such lexical pattern might be used as templates for creating the axioms for those related classes. If a lexical pattern is found in the labels of external ontologies, this might indicate that some content and axioms might be re-used and shared. For instance, the lexical pattern *virus* is not a class despite it is a relevant concept in

an ontology about vaccines. Besides, *virus* has been found as the full label of classes in another 4 BioPortal ontologies. This fact does not mean that these concepts are equivalent, but at least they should be considered by a domain expert to be re-used when the need for a 'virus' class arises. Something similar oc-curs with the lexical pattern *canine* which is not a class but could be defined and linked with labels that contain it to indicate that these vaccines are focused on ca-nines. We cannot only re-use classes but also properties. For instance, the lexical pattern *encoding* of the Vaccine Ontology is an object property of the Bone Dysplasia Ontology (http://purl.org/skeletome/bonedysplasia#encoding).

Our method represents the groups of tokens found in the labels of an ontology as a graph, which is built as the ontology is processed. For each class we extract its label, which is split using blank as delimiter. After this, each token is a node of the graph, and each arrow represents that the tokens appear consecutively and in that order in a label. We also store additional information for each node like the position in the label (e.g., the same word could appear several times in the same label) and the URI of the class to speed up further analyses. Once the ontology labels are represented in the graph, we apply our algorithm to identify lexical patterns within the ontology. Finding a lexical pattern of length N requires to navigate through N edges starting from an initial node. We can obtain the whole set of lexical patterns within an ontology by repeating the process in all the nodes of the graph. Despite we considered options like n-gram or high-performance graph databases, the complexity of the links between our tokens and the need for a fast answer given the size of many biomedical ontologies led us to develop this graph representation.

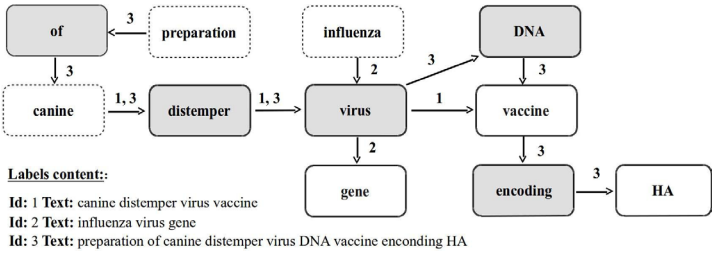

Fig. 1. Graph representation of the content of three labels

Figure 1 shows the representation of three labels using our representation. Underlined boxes represent initial nodes; solid white boxes stand for the final node of labels; and solid gray boxes are neither initial nor final nodes. In any case, a node can play different roles at the same time such as *canine*, which is an initial and intermediate node. The arrows represent edges and the numbers refer to

the identification of the labels where it appears. If we analyze this figure, the pattern *canine distemper virus* (length 3) can be found. The node *virus* has input arrows from labels 1, 2 and 3. This means that it is contained in these three labels. The arrow between *distemper* and *virus* is labelled 1,3, and this means that both words appear consecutively in these two labels being a lexical pattern of length two and with two repetitions. Consequently, *distemper virus* is a lexical pattern that appears in labels 1, 3 but not in 2.

Once the graph is built, we filter out some groups of words. On the one hand, groups of words which consist only of stop-words, that is, words without meaning, are filtered. Every word group must reach a *coverage* threshold, which is the minimum percentage of labels in which a group of words must appear to be considered a lexical pattern in a particular analysis process. This enables different analyses demands to be performed and to adjust the result set to different goals of the ontology designer. Our method does not need to rebuild the graph for each change in the coverage threshold.

2.2 Shared Lexical Patterns

Once the set of lexical patterns for a given ontology has been obtained, we could look for such patterns in the labels of external ones. In our context, this means that it should be likely to find common lexical patterns in different ontologies. However, finding matches would not mean that the corresponding classes are equivalent or refer to the same domain knowledge, though the inspection of such alignments could be interesting for the designer. In our method, we identify an *exact match* when a lexical pattern and the label of an external ontology are the same. It is worth pointing out that the lexical patterns could or not being a class in the input ontology. Finding such shared lexical patterns and, even more importantly, such shared classes is relevant for the enrichment of the ontologies since existing axioms in external ontologies could be re-used in the source ontology.

Furthermore, we propose the inspection of IRIs for the *exact matches* to check if they refer to the same concepts. For instance, the lexical pattern *protein* is found as a class with the same IRI http://purl.obolibrary.org/obo/CHEBI_36080 in *CHEBI*, *microRNA Ontology*, *Gene Expression Ontology* and *Regulation of Gene Expression Ontology*. That might be a sign of good re-use. In other cases we could not assume this fact. For instance, the lexical pattern *influenza* has been found as a class in 7 external ontologies with 7 different IRIs. Knowing if they refer to seven different concepts is beyond the scope of this work, but it might be a sign of a lack of re-use.

3 Results

We have analysed the BioPortal ontologies publicly accessible in OWL and OBO format in December 2012. The analysis has been supported by a home-made software tool called OntoEnrich [5]. Since this tool only works with OWL ontologies, we used the OWL Syntax Converter (http://owl.cs.manchester.ac.uk/ converter/) to get OWL versions of the ontologies only available in OBO for-mat. In this way, our base contained 286 ontologies (177 OWL, 109 OBO). Given that 44 had no labels, we automatically generated the labels from the IRIs. 70 ontologies were not processed because of importing inaccessible OWL files or failure in the OBO to OWL conversion.

We have analyzed the labels of the ontologies for different coverage values: from 1% to 5% with increments of 1. Here we show the results with the coverage set to 1%, but the complete set is available at http:\miuras.inf.um.es/aime. For each ontology, we obtained a series of metrics, the main ones being:

- *Number of labels*: number of labels in an ontology.
  - *Number of lexical patterns*: number of lexical patterns found in the ontology.
  - *Classes affected by lexical patterns*: number and percentage of classes in which lexical patterns are found.
  - *Classes affected by matches*: number and percentage of classes for which exact matches are found.
  - *Repetition of words*: percentage of repeated words in the ontology labels.

3.1 Global Characterization of BioPortal Ontologies

Table 1 shows the summary of the characterization of the 216 ontologies analyzed, whose main findings are described next.

- Labels: 90% of the classes have labels. This value suggests that the BioPortal ontologies are rich in labels, so their analysis might be interesting. Besides, 68.5% of the words used in labels are repeated and this is a sign of regularity. The mean of repeated labels is 0.90% which is a good result, but the maximum value of 31%, which means that almost 1 out of 3 classes share a full label. This ratio is greater than 5% for 7 ontologies.
- Lexical patterns: With a coverage of 1%, the mean number of lexical patterns is 63, the highest being 555 repetitions in the *Ontology of Data Mining*. The mean percentage of classes for which patterns are found is 56%. This means that many BioPortal ontologies may have regularities in their labels.
- Matches in external ontologies: The mean number of external matches per ontology is 124.7. This means that many patterns of each ontology are found many times in other BioPortal ontologies as full labels of classes. The mean number of external matches per lexical pattern is 44% and the percentage of

classes that are covered by lexical patterns with external matches is 46% so these classes contain knowledge that exists in other resources, making the possibility of reusing content from external ontologies evident.

3.2 Cluster Analysis

We have applied agglomerative hierarchical clustering to the percentage of classes with patterns with coverage 1%. The analysis of the dendrogram suggested the existence of three main groups of ontologies. Then, we applied k-means (k=3) to get a representation of the three groups and to get each ontology associated with one of the three clusters obtained, whose centroids have the values shown in Table 2.

Table 1. Summary of data for different analysed variables

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Number of classes		3	132	339	5678	1544	286400
Number of labels		1	98	290	5130	1285	232300
Repeated Words in labels	%	50.00	61.32	68.25	68.50	76.05	97.59
Repeated labels	%	0	0	0	0.9085	0.2740	31.86
Number of Lexical Patterns		0	10	42	63.6	85	555
Classes covered by LPs	%	0	36.62	63.33	56.67	80.80	100
Number of external matches per ontology		0	18	68	124.7	122	1298
Number of external matches per LP	%	0	25.29	45.05	44.47	64.10	100,00
Classes covered by LPs with external matches	%	0	20	46.02	43.66	67.19	100

Table 2. Centroid of the clusters

Variable	Cluster1	Cluster2	Cluster3
%ClassesWithPatterns	82.25	48.84	7.87
Ontologies	87	76	53

The cluster analysis splits the base of ontologies in three differentiated groups. Cluster 1 includes the ontologies for which most of the classes of the ontology have lexical patterns associated. The ontologies of this cluster are the most suitable for applying enrichment methods by exploiting the lexical patterns, and they represent 40% of the ontologies analyzed. Cluster 2 includes the ontologies for which around half of the classes of the ontology have associated lexical patterns. We cannot say these ontologies could not benefit from the application of enrichment methods, but they could be assigned a lower priority. This cluster includes 35% of the ontologies, which means that 75% of the BioPortal ontologies analyzed could benefit from enrichment processes. Finally,

Cluster 3 includes those ontologies whose classes do not have many patterns associated. Consequently, they seem to be not very interesting for the enrichment based on the exploitation of the lexical patterns. This group has 25% of the ontologies analyzed. This cluster also includes the ontologies for which patterns have not been found with coverage 1%. The members of each cluster are listed at http:\miuras.inf.um.es/aime.

3.3 Cluster Analysis of the OBO Foundry Ontologies

As mentioned in the introduction, the OBO Foundry defined a series of principles for building biomedical ontologies, among which using rich labels and a systematic naming convention are relevant for this work. If ontology builders follow such principles, the structure and content of the labels should be descriptive about the meaning of the concept. We have analysed which OBO Foundry ontologies are associated with each cluster. It should be noted that, at the time of writing, there are not many ontologies recognized as OBO Foundry members, but more that are candidate to be OBO Foundry ontologies. For an ontology to be a member, the OBO Foundry must have checked that they have been developed by following different criteria.

The OBO Foundry member ontologies are: Gene Ontology, CHEBI, Phenotypic quality, Protein Ontology, Xenopus anatomy and development and Zebrafish anatomy and development. The Gene Ontology appears in Cluster1, CHEBI in Cluster3, the Phenotypic quality ontology could not be processed and the other three ontologies appear in Cluster2. These results are in line with our expectations. Given that CHEBI is oriented to types of chemical entities and given its size it is likely that the patterns found do not appear in at least 1% of the classes. The results of using the lowest coverage (around 0%) show 16 out of the most frequent 20 lexical patterns have a frequency below 1%.

The analysis of the candidate ontologies reveals that 33 ontologies belong to Cluster1, 32 to Cluster2 and 10 to Cluster3. This means that 86% of the ontologies analysed are in Cluster1 or 2, what can be interpreted as these ontologies follow the guideline for labels.

3.4 Analysis of Lexical Patterns

Table 3 describes the set of patterns obtained with 1% of coverage considering the whole set of analyzed ontologies. It should be pointed out that 31% (4011) of the patterns appear in more than one ontology. We have analyzed the impact of using coverages 2%, 3%, 4% and 5% in the results obtained. Figure 2 shows the percentage of lexical patterns found for each coverage value, considering the total number of patterns found using such coverages. It can be seen that the

5% of coverage hardly ever find lexical patterns (mean= 8.92 patterns) and that many ontologies follow similar distributions of percentage of lexical patterns.

Table 3. Numerical metrics about the precise analysis of the lexical patterns

Lexical Patterns		Length			Frequency		
Total	**Unique**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**
13805	9494	1,0	1,841	12,0	2,0	115,6	5660,0

The coverage is a frequency threshold, so increasing its value means that the least frequent patterns would be removed. We have also analyzed the impact of the coverage in the distribution of ontologies after applying the clustering. The expected result was a reduction in ontologies in Cluster1 since we are removing patterns and, therefore, less classes are affected. The number of ontologies in Cluster2 remains relatively stable, while the number of ontologies drops from 87 to 28 in Cluster1. However, the number of ontologies in each group is quite similar for 3%, 4% and 5%. This means that many ontologies present many regularities with frequency lower than 3%. The same result has been found for the OBO Foundry ontologies. Two member ontologies move from Cluster2 to Cluster3 with coverage 5%. For such coverage, the distribution of ontologies is: Cluster1 (8), Cluster2 (32) and Cluster3 (35).

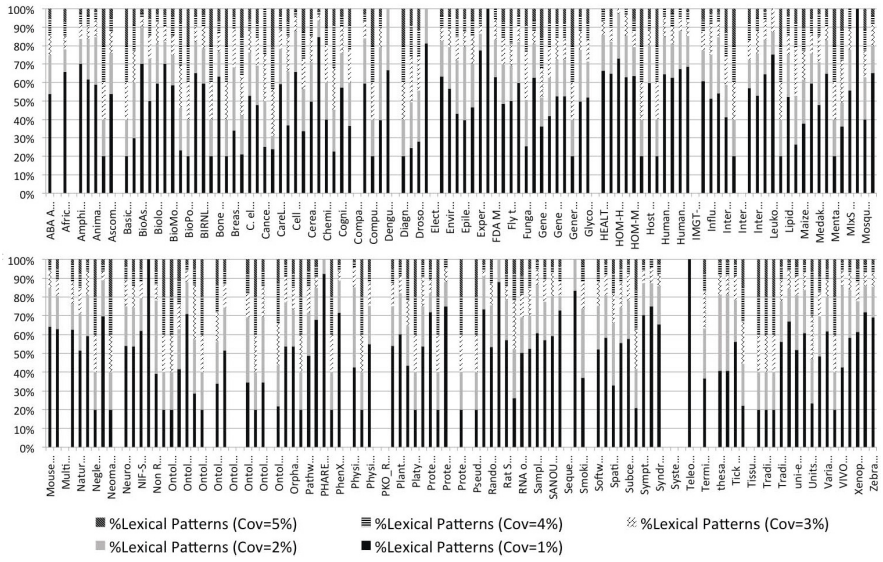

Fig. 2. Percentage of lexical patterns found for different coverages (1%, 2%, 3%, 4% and 5%), considering the total number of lexical patterns found using such coverages

3.5 Analysis of Re-use of Concepts

We have studied the external matches found in terms of reusability. As mentioned, if both source and target classes of an external match for a given pattern share the same IRI, this would be a sign of good re-use. We have then analysed whether such IRI sharing happened in the external matches. The mean value of external matches is 3.357 and the number of IRIs is 2.651, so around 79% of the source and target classes have different IRIs. For instance, the lexical pattern *human* appears in 5 ontologies using 5 different IRIs, and *infection* appears in 9 ontologies using 9 IRIs too. We have not inspected whether the corresponding classes have equivalence axioms linking the different IRIs.

4 Conclusions and Further Work

Important challenges in healthcare, like the achievement of semantic interoperability of healthcare records, require the use of good, useful ontologies for different purposes. Bioportal ontologies contain many ontologies rich in text content but not so rich in axiomatic content. The axiomatic enrichment of such ontologies could be done by exploiting the content and structure of the labels. In this paper, we have analysed the labels of Bioportal ontologies and we have been able to classify them in terms of suitability for applying enrichment processes. From our results, we suspect that re-use is not used by the biomedical ontologies builders as much as they should, although we should perform a more detailed analysis of the external matches. We think we could develop an ontology-dependent metric to get an appropriate coverage threshold. The regularities in labels of some ontologies might happen in a particular area of the ontology. This might be the case of highly modularized ontologies with several independent modules, in which regularities in the labels within each module. Those patterns might be filtered out with a coverage of 1%, but our method and our OntoEnrich tool permit to adjust the coverage to the user needs. In this work, the re-use of concepts has been analyzed from a lexical perspective. Bioportal stores mappings between the Bioportal ontologies and we plan to compare such mappings with our external matches to improve our knowledge about re-use in biomedical ontologies. As a result of this work, we know more about the lexical properties of Bioportal ontologies, which will permit us to develop effective axiomatic enrichment processes.

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

References

1. Consortium, G.O.: Gene Ontology: tool for the unification of biology. Nature Genetics 23, 25–29 (2000)
2. European Commission. Semantic interoperability for better health and safer health-care. deployment and research roadmap for Europe (2009) ISBN-13 : 978-92-79-11139-6
3. Fernandez-Breis, J.T., Iannone, L., Palmisano, I., Rector, A.L., Stevens, R.: Enriching the gene ontology via the dissection of labels using the ontology pre-processor language. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS, vol. 6317, pp. 59–73. Springer, Heidelberg (2010)
4. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.-A.D., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37(Web-Server-Issue), 170–173 (2009)
5. Quesada-Martínez, M., Fernández-Breis, J.T., Stevens, R.: 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 Do-main (K-MED 2012), Galway, Ireland, pp. 1–10 (2012)
6. Quesada-Martínez, M., Fernández-Breis, J.T., Stevens, R.: 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 2012, pp. 7–16. ACM, New York (2012)
7. Third, A.: "Hidden semantics": what can we learn from the names in an ontology? In: 7th International Conference on Natural Language Generation (2012)