Prioritising Lexical Patterns to Increase Axiomatisation in Biomedical Ontologies

The Role of Localisation and Modularity

M. Quesada-Martínez; J. T. Fernández-Breis; R. Stevens; E. Mikroyannidi

1. Introduction

In recent years, the medical informatics community has made efforts to achieve semantic interoperability in healthcare. The Semantic Health project defined a

Methods: We propose metrics for suggesting which lexical regularities should be the starting point to enrich complex ontologies.

Conclusions: The locality and distribution of lexical patterns offer insights into the further engineering of the ontology. Developers can use this information to improve the axiomatisation of their ontologies.

Keywords
Biological ontologies, ontology enrichment, quality assurance, lexical patterns

Summary
Introduction: This article is part of the Focus Theme of Methods of Information and Medicine on “Managing Interoperability and Complexity in Health Systems”.

Objectives: In previous work, we have defined methods for the extraction of lexical patterns from labels as an initial step towards semi-automatic ontology enrichment methods. Our previous findings revealed that many biomedical ontologies could benefit from enrichment methods using lexical patterns as a starting point. Here, we aim to identify which lexical patterns are appropriate for ontology enrichment, driving its analysis by metrics to prioritise the patterns.

Methods: We propose metrics for suggesting which lexical regularities should be the starting point to enrich complex ontologies.

Results: The metrics provide information about the engineering of the ontologies and the relevance of the patterns. Our method enables the suggestion of links between classes that are not made explicit in the ontology. We propose a prioritisation of the lexical patterns found in the analysed ontologies.

Conclusions: The locality and distribution of lexical patterns offer insights into the further engineering of the ontology. Developers can use this information to improve the axiomatisation of their ontologies.

Correspondence to:
Manuel Quesada-Martínez
Universidad de Murcia
Departamento de Informática y Sistemas
Facultad de Informática Campus de Espinardo
30100 Murcia
Spain
E-mail: manuel.quesada@um.es

Methods Inf Med 2015; 54: 56–64
10.3414/ME13-02-0026
received: June 17, 2013
accepted: May 7, 2014
epub ahead of print: July 4, 2014

http://dx.doi.org/10.3414/ME13-02-0026

Quality assurance and ontology enrichment have been addressed in several ways and they require the combination of different activities at both textual and axiomatic levels. On the one hand, ontology learning techniques [6] have been used for automating the development and maintenance of biomedical ontologies [7]. Ontology learning approaches [8] can be subdivided by extraction tasks: terms, synonyms, concepts, relations, and axioms. For instance, the ODIE project uses Natural Language
Processing (NLP) techniques to identify and retrieve relevant free text information from clinical document repositories using ontological terminology, with the goal of improving and enriching ontologies. Many such approaches are focused on the automatic construction of the ontology. In contrast, ontology enrichment starts from a given ontology and has the aim of generating additional concepts or axioms using statistical data about the usage of the name of the concepts of the ontology in a text corpus [6]. The NLP requirements of ontology enrichment are also related to the interpretation of text [9, 10]. Lexico-syntactic patterns [11, 12] detect in sentences like "systemic granulomatous diseases such as Crohn’s disease ..." that "Crohn’s disease" is a type of "systemic granulomatous diseases", and the exploitation of compound or multi-word terms may help to identify hierarchical relationships. From "prostatic carcinoma" we can infer that it is a kind of "carcinoma". Statistical approaches are based on Firth’s notion; “a word is characterised by the company it keeps” so the analysis of co-occurrences of a word play an important role in the classification of such a word [7].

Some enrichment methods exploit the hidden semantics embedded in the labels of the ontologies, but are not made explicit through axioms [5]. The Gene Ontology (GO) [13] has been the target of most enrichment efforts [14, 15], although other generic methods, not ad-hoc to a particular ontology, would be desirable. We have proposed an approach for ontology enrichment based on the detection and analysis of such pattern types, called Lexical Patterns (LP), which are the result of applying a statistical approach of NLP using the labels of ontology classes as its corpus. For instance, the GO has 40043 classes. An example of an LP is "regulation of", which appears in 20.85% of the classes, so this LP can be used as a seed to identify lexico-syntactic patterns of the type of "regulation of X" where X is codified as the label of a class. These patterns may then be shown to the ontology developer, who could add new axioms to the ontology. Our results in the extraction of LPs from BioPortal ontologies [16] show that the exploitation of such patterns has potential for gaining insights about the further engineering and enrichment of the ontologies. However, since biomedical ontologies have many LPs [16], detecting the most relevant ones for enrichment should make this task easier for the ontology developer.

On the axiomatic side, the manual detection of irregularities like missing restrictions has been addressed in work such as [17, 18], whereas syntactic and semantic irregularities are detected by RIO [19]. In addition, tools like OOPS are able to detect pitfalls in the axiomatisation of ontologies [20]. Hence, including methods that pinpoint anomalies would also help ontology developers to enrich their ontologies. The use of metrics is common practice in engineering activities and this also happens in ontology engineering. Metrics are widely used to evaluate ontology quality, correctness or similarity [21–24].

The main objective of this current work is to extend our method for the analysis of LPs by adding metrics to study where LPs occur and how they are distributed in the ontologies. Our hypothesis is that by analysing the labels of an ontology to yield LPs and the locality of those patterns, then these will give information to drive the further axiomatisation of that ontology. For instance, the LP 'sebaceous' (the DOID ontology) appears in labels of descendant classes of 'disease of cellular proliferation' like 'sebaceous adenoma'; 'sebaceous adenocarcinoma' or 'sebaceous gland neoplasm'. Moreover, it appears just once in 'sebaceous gland disease', which is a descendant of 'disease of anatomical entity'. However, there is no link between both classes.

Two types of metrics are addressed in this work:

1) distance between the classes that contain a lexical pattern; 2) distribution of the lexical pattern in a particular context in the ontology.

These metrics do not only contribute to the improvement of ontology enrichment methods, but are also relevant for quality assurance, because the lexical findings may be suggestive of anomalies in the axiomatic content of the ontologies.

2. Methods

2.1 Detecting Lexical Patterns in Labels

Our concept of Lexical Pattern was introduced in [16, 25] and captures the assumption that repeated groups of words are likely to encode some domain meaning.

- **Definition 1 (Lexical Pattern, LP):** an LP is a list of tokens that are repeated in different labels of an ontology. Given an ontology \( \theta \) with a set of classes \( OCS = \{OC_1 \ldots OC_i\} \) an LP is a set of tokens repeated in the same order and sequence in different labels associated with the elements of OCS.

- **Definition 2 (Reduced Ontology Class Set, ROCS):** given an LP, it has associated a Reduced Ontology Class Set (ROCS) with those classes that exhibit the LP: \( ROCS \subseteq OCS \).

The content of the labels of the ROCS classes may provide implicit information about the concept participating in the process described by the LP. For instance, "regulation of pigmentation during development" is in the ROCS of "regulation of", and "pigmentation during development" is a class in the GO (Figure 1). Hence, axioms linking such entities would enrich the ontology. Let us consider the LP of the Gene Ontology ‘regulation of’ that appears in labels such as ‘positive regulation of mRNA’ and ‘negative regulation of mRNA’. These two labels can be decomposed as:

\[
\text{<type of regulation; the action of regulation; substance regulated>}
\]

Our LP extraction method [16] represents labels of the ontology as a graph of tokens that is built as the ontology is processed. Each node of the graph corresponds to one word and each arrow means that the nodes so linked appear consecutively and in that order in a label. We also store additional information for each node like the position in the label or the URI of the class. This representation allows the frequency of an LP to be found and it offers the possibility of filtering out those patterns whose frequency is not considered high enough by the ontology developer. Such a minimum
frequency is called the Coverage Threshold.

- **Definition 3 (Coverage threshold):** the coverage threshold of an LP is the minimum percentage of classes (out of the total number) in which such LP must appear to be accepted.

### 2.2 Locating Lexical Patterns within the Class Hierarchy

The classes exhibiting LPs can be distributed across the ontology hierarchy. Measuring the closeness of the classes, within the hierarchy, associated with an LP can provide some insights about the engineering of the ontology. The hypothesis is that similar patterns will occur in classes that are located close to each other in the hierarchy. Likewise, the existence of similar patterns that are far apart may suggest a defect – a "bad term smell". The localisation of the set of classes exhibiting an LP might help ontology developers to understand each LP and prioritising those that can be used to systematically create axioms from the classes that exhibit each LP. Different situations and examples are discussed later in this paper.

Our definition of locality is based on semantic similarity measures, which, according to [24], provide a numeric score for two ontology classes that reflect their closeness in meaning. There are two approaches for calculating the semantic similarity of two ontology classes: edge-based and node-based [24]. On the one hand, edge-based approaches count the number of edges in the graph path between two classes. [26] proposed one of the first edge-based structural approaches measuring path distances between classes. These approaches assume that all of the semantic links are equally weighted but, generally, the greater distance from the root is, the more specific the classes are. On the other hand, node-based approaches also rely on comparing the properties of the classes involved [24] (e.g., measuring the probability of occurrence of classes in a specific corpus [27]). Semantic similarity measures have also been used for calculating functional similarity between gene products, which are annotated with classes from the Gene Ontology [24, 27, 28].

In this paper, we use the distance function ($DF$) between ontology classes used in [29], and that combines edge-based and graph-based approaches. It measures the hierarchical distance between two classes $OC_i$ and $OC_j$ using two sub-graphs extracted from both classes. Each subgraph ($cA(OC_n)$) contains those classes (ancestors) in the path from a class $OC_m$ to the root through "subClassOf" relationships. The intersection of two subgraphs represents their set of common ancestors, which is denoted as $int$. Thus, the distance between two classes, $DF(OC_i, OC_j) \in [0, 1]$, and it is calculated as:

$$DF(OC_i, OC_j) = \frac{\text{int}(cA(OC_i), cA(OC_j))}{\text{max}(\text{size}(cA(OC_i)), \text{size}(cA(OC_j))))}$$

The score is high in case of close classes. ▶ Figure 2 shows two partial extractions of the hierarchy of classes in the Human Disease Ontology (DOID). The highlighted class ‘central corneal ulcer’ has taxonomic depth 11 and 10 ancestors: ‘corneal ulcer’, ‘keratitis’, ‘corneal disease’, etc. The classes ‘central corneal ulcer’ and ‘perforated corneal ulcer’ are siblings, so $DF$ is 1. The distance between ‘posterior uveal melanoma’ ($OC_i$) and ‘posterior pituitary gland neo-
plasm" (OC₂) is 0.78 (|CA(OC₂)| = 9, |CA(OC₁)| = 10, int = 4) because they are in different subtaxonomies.

We have adapted our distance function to measure the locality of an LP. LPs and classes are related through the ROCS (e.g., the set of classes that exhibit an LP). The highlighted classes of Figure 2 left are in the ROCS of the LP 'corneal ulcer'. In this case, ontology authors have followed a systematic naming in the class labeling that contributes to the lexical similarity of the ROCS classes. The class 'corneal ulcer', which matches the lexical pattern, is the ancestor of the other ROCS classes, so the hierarchical relationships are already explicit. However, the highlighted classes in Figure 2 right are in the ROCS of the LP 'posterior'; these classes do not appear in the same hierarchy but the LP encodes some meaning related to the physical position that should be represented in the axioms of those classes.

Definition 4 (locality measure of a lexical pattern). Given the ROCS of an LP, the locality measure of this LP is the median of the distance function over all the pairs in the cartesian product ROCS × ROCS. This locality measure is in the range [0, 1]; 1 means close in the hierarchy, and 0 far away.

The locality measure of 'corneal ulcer' is 0.85 and the locality measure of 'posterior' is 0.14. In this latter case, this happens because it appears as a modifier in many labels that are not grouped in the hierarchy e.g., 'posterior pituitary gland neoplasm' and 'posterior uveal melanoma' which are in distinct sub-branches 'mucouskeletal system cancer' and 'nervous system cancer'.

2.3 Grouping of Lexical Patterns

The locality measure provides general information about the distribution of the LPs across the ontology. Some ontologies are structured inclear, if implicit, modules. For instance, the Human Disease Ontology has subtaxonomies of diseases by infectious agent, anatomical entity, cellular proliferation, etc. Likewise, the three aspects of the GO, namely, Molecular Function, Biological Process and Cellular Component are defined as three independent taxonomies.

Thus, detecting the modularisation of an LP in a particular context of the ontology also provides interesting information about the engineering of the ontology.

Definition 5 (modularisation measure of a lexical pattern). The modularisation of an LP is the distribution of the classes of its ROCS in a certain context of the ontologies defined by a set of classes of interest. Formally, given an ontology θ, an LP exhibited in ROCS = {OC₁ ... OCₖ}, the modularisation calculates for each class of interest OCₖ the percentage of classes from the ROCS that have OCₖ as ancestor.

In our running example of the GO, 'kinase activity' is an LP that appears in 0.74% of the classes. Some classes in its ROCS are 'kinase activity', 'regulation of kinase activity', 'dolichol kinase activity', and 'protein histidine kinase activity'. We calculate the modularisation of this LP using as classes of interest the main classes of the three GO aspects: 'biological_process’ (BP), 'cellular_component’ (CC) and ‘molecular_function’ (MF). This LP appears in classes that are descendants of BP and MF, and corresponds to the full label of a class in MF. Moreover, in BP, it is preceded by another LP, 'regulation of', as in 'regulation of kinase activity'. As a consequence, the presence of this LP in two classes of interest could help the ontology developer to make explicit the links between the corresponding MF and BP classes.

LPs may have a different number of occurrences in a different number of modules, so the normalised standard deviation (SD) for each pattern is between 0 and 1. The deviation of patterns appearing in one
module of interest will be close to 0, and those whose distribution is less homogeneous will be closer to 1. We classify LPs in three categories: (G1) LPs with $SD = 0.0$ that are localised in the same module; (G2) co-LPs that appear in different modules and are homogenously dispersed, so they need to be inspected in advance, $SD = (0.0, 0.4)$; (G3) LPs that appear mostly in one module, $SD = (0.4, 1)$.

We can now revisit the example of the LP 'sebaceous' (the Human Disease Ontology) used in the introduction. By adding an axiom like 'sebaceous adenoma subClassOf sebaceous gland disease', automated reasoning could infer that 'sebaceous adenoma sub-ClassOf (disease of cellular proliferation) and (disease of anatomical entity)'. In this case, finding where the LP appears could help an ontology developer to identify the need for such axiom. Finally, the LPs can be prioritised for their axiomatic enhancement using the locality and modularity metrics.

## 3. Results

Here we analyse the main results, which will be further discussed in Section 4. The complete results can be found at http://miuras.inf.um.es/mixhs/SI2013.

### 3.1 Materials

We have analysed four biomedical ontologies, which were selected due to their size and different content: 1) the Human Disease Ontology (DOID); 2) the Chemical Entities of Biological Interest (ChEBI); 3) the Gene Ontology (GO); and 4) SNOMED CT. These experiments have been run using our OntoEnrich tool [25], which implements the lexical analysis methods, and R (http://www.r-project.org/) for the data analysis.

### 3.2 Description of the Lexical Patterns Found

The set of LPs are quantitatively described in Table 1 using the coverage thresholds 0.1, 0.5 and 1.0. The number of patterns ranges from less than ten to thousands in the case of the GO.

### 3.3. Localisation of the Lexical Patterns in the Ontology

A graphical summary of the locality measure of the LPs for each coverage threshold is shown in Figure 3. The mean value of the locality measure ranges from 0.2 to 0.48, which means that, on average, LPs are distributed along the hierarchy.

![Graphical summary of the locality measure of the LPs](image)

### Table 1

Quantitative description of the LPs obtained for the four ontologies of our corpus. The first two columns describe the ontology indicating its name and the number of classes. After that, there are three groups of two columns; every group is labelled with 'LPs Cov. = X' where X is the coverage threshold used to extract the LPs for each ontology. For each combination (Ontology Name, Cov) the table shows three columns: (1) ‘Num LPs’ or number of LPs found for such configuration, (2) ‘% LPs Classes’ or the percentage of LPs that are found as the whole label of a class, and (3) ‘Mean ROCS Size’.

<table>
<thead>
<tr>
<th>Ontology Name</th>
<th>Num. Classes</th>
<th>Coverage = 1.0</th>
<th>Coverage = 0.5</th>
<th>Coverage = 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Num. LPs</td>
<td>% LPs Classes</td>
<td>Mean ROCS Size</td>
</tr>
<tr>
<td>CHEBI</td>
<td>39397</td>
<td>4</td>
<td>50.00</td>
<td>1504.3</td>
</tr>
<tr>
<td>Human Disease</td>
<td>6311</td>
<td>25</td>
<td>32.00</td>
<td>164.8</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>37381</td>
<td>75</td>
<td>21.33</td>
<td>1537.7</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>295481</td>
<td>43</td>
<td>0.00</td>
<td>5616.8</td>
</tr>
</tbody>
</table>

For personal or educational use only. No other uses without permission. All rights reserved.
3.4 Modular Distribution of the Lexical Patterns

We have calculated the modularisation measure for each LP using the classes at the lowest taxonomic depth with more than one class as classes of interest. In the case of the ChEBI and the GO, the direct descendant classes of the root (level 1) constitute the classes of interest. In the case of the DOID, there is just one class that is a direct descendant of the root (‘disease’) so we chose its 8 direct descendants (level 2) as classes of interest. This includes classes like ‘disease by infectious agent’, ‘disease of anatomical entity’, ‘genetic disease’, ‘medical disorder’, etc. Likewise, we have selected elements from level 2 in SNOMED CT. The quantitative description of each module is: 1) the DOID, 8 classes found in level 2, 2) the ChEBI, 3 classes found in level 1, 3) the GO, 3 classes found in level 1, and SNOMED CT, 19 classes found in level 2.

A graphical summary of the modularity metric for each ontology and coverage threshold is shown in Figure 4. The mean percentage of classes of interest in which an LP appears is 52.79% (1.56) in the ChEBI, 31.62% (2.48) in the DOID, 67.06% (2.0) in the GO, and 28.07% (5.32) in SNOMED CT (the figures in brackets are the absolute values, for instance, LPs from the GO appear as descendant of 1.56 classes of interest).

Finally, we show the distribution in categories of the LPs, according to the standard deviation value after calculating the modularisation measure, as a tuple <G1, G2, G3>. For instance, in the ChEBI <41.62, 0.00, 58.38> means that 41.62% of the patterns are in G1, 0% in G2 and 58.38 in G3. The values for the other ontologies are the DOID <29.73, 28.45, 41.82>, the GO <32.55, 12.78, 54.66> and SNOMED CT <44.19, 48.82, 6.99>.

3.5 Prioritisation of the Lexical Patterns

In Table 2 we show the prioritisation of the LPs (obtained with a 0.1% of coverage) with highest values of the standard deviation of the modularisation measure and that appear in different parts of the ontology according to the context defined by the classes of interest (see sub-section 3.4). These patterns can be used to create axioms as, we explained in Section 2.1.

4. Discussion

4.1 Lexical Patterns

The number of LPs found is not directly proportional to the number of classes, since larger ontologies may have a lower number of patterns, regardless of the value of the coverage threshold. For instance, the number of classes of the ChEBI and the GO are similar, but on average the number of LPs is 11.75 times larger in the GO. The number of LPs found is not as systematic as in the GO due to the nature of the domain content, so the number of LPs is lower. Moreover, the ChEBI includes terms that are not considered in our current analysis like ‘3-alpha-D-glucuronosyl-2-palmitoyl-[(10R)-10-methyloctadecanoyl]-sn-glycerol’ as we take a blank character as a delimiter in finding tokens. However, this is not a

Figure 4 Graphical summary of the modularisation measure of the LPs after analysing the four ontologies at the three coverage thresholds (‘Cov = 1.0’, ‘Cov = 0.5’, ‘Cov = 0.1’)
limitation of the approach, but of our current implementation.

Finally, LPs obtained from the DOID or SNOMED CT do not identify processes or roles like in the GO but refer mainly to: 1) general concepts that should be defined as classes in upper-levels of the hierarchy such as: ‘carcinoma’, ‘syndrome’, etc; 2) modifiers indicating properties of the classes like: ‘acute’, ‘benign’, ‘childhood’, ‘adult’, etc. In the case of SNOMED CT, some patterns like ‘structure of…’, ‘… due to…’, ‘neoplasms of…’, ‘finding of…’ could be explored as they could be defined as the composition of other concepts.

4.2 Localisation and Modularisation

We have presented two metrics that quantitatively locate the patterns in the hierarchy of classes and aim to help in the prioritisation of the patterns which would permit ontology developers to focus on the most relevant first: 1) the locality measure shows how LPs are distributed across the ontology hierarchy; 2) the lexical modularity shows how LPs are distributed in the class hierarchy with respect to a context of the ontology.

The mean value of the locality measure means that, on average, LPs are distributed along the hierarchy. When a systematic naming convention is applied, the label of a class includes part of the label of its direct ancestor. The common fragment is identified in our approach as an LP, but the locality metric decreases when the contains classes at very different taxonomic depth. This is why locality has to be complemented with modularity, which could reveal whether the elements of are related to the same subhierarchies, as happens with the example of the regulation of the kinase activity.

In terms of modularisation, the highest relative values for the ChEBI and the GO are due to finding similar lexical entities in different branches of the hierarchy considering the classes of interest as branches. This means that there should be axioms between entities of the different branches, between classes of the different GO aspects in the case of the GO. Lower values are due to the existence of more classes in the upper level of the hierarchy, and that classes of those branches have fewer connections with classes from other ones.

We use state-of-the-art distance functions to measure the distance between two classes. Given that LPs are associated with sets of classes in the ontology and that the locality of an LP depends on the classes included in their ROCS, the use of distances between classes is appropriate. This process requires calculating the distances of many pairs of classes, so the final locality of a given LP requires aggregating the values for such pairs. We decided to use the median because it is more robust and less prone to arbitrary results or to be affected by contamination than other measures like the mean or the mode. For instance, the mean of a ROCS where the values of distance functions for its pairs are \{0.13, 0.15, 0.13, 0.14, 0.13, 0.14, 0.17, 0.86, and 0.75\} is 0.27, whereas the median is 0.14. However, locality is not enough when the ontology

<table>
<thead>
<tr>
<th>ROCS Size / SD of MM of LPs</th>
<th>CHEBI</th>
<th>ROCS Size / SD of MM of LPs</th>
<th>Human Disease</th>
<th>ROCS Size / SD of MM of LPs</th>
<th>Gene Ontology</th>
<th>ROCS Size / SD of MM of LPs</th>
<th>SNOMED CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>482 / 0.49896</td>
<td>B</td>
<td>427 / 0.49764</td>
<td>carcinoma</td>
<td>2054 / 0.49975</td>
<td>biosynthetic process</td>
<td>600 / 0.49907</td>
<td>Deficiency of</td>
</tr>
<tr>
<td>646 / 0.49844</td>
<td>A</td>
<td>146 / 0.49655</td>
<td>cell carcinoma</td>
<td>1422 / 0.49964</td>
<td>catabolic process</td>
<td>511 / 0.49545</td>
<td>caused by</td>
</tr>
<tr>
<td>299 / 0.49832</td>
<td>phosphate</td>
<td>83 / 0.49390</td>
<td>squamous</td>
<td>807 / 0.49938</td>
<td>differentiation</td>
<td>348 / 0.48922</td>
<td>occupant of</td>
</tr>
</tbody>
</table>
| 3726 / 0.49797             | acid | 185 / 0.49189              | Tumor          | 628 / 0.49920              | formation       | 666 / 0.48144 | accident involv-
| 1163 / 0.49784             | froup | 56 / 0.49107               | lymphoma       | 296 / 0.49830              | interaction     | 319 / 0.47692 | to use       |
| 195 / 0.49742              | derivative | 195 / 0.48974       | neoplasm       | 278 / 0.49819              | activation      | 1417 / 0.46881 | Blood group  |
| 163 / 0.49691              | diphosphate | 87 / 0.48850           | sarcoma        | 267 / 0.49812              | signal transduction | 471 / 0.46781 | monitoring   |
| 154 / 0.49673              | molecular | 40 / 0.48750          | benign neoplasm| 2472 / 0.49776             | positive regulation | 321 / 0.46597 | Motor vehicle |
| 139 / 0.49637              | compound | 37 / 0.48611           | mucinous       | 370 / 0.49728              | receptor signaling pathway | 536 / 0.46543 | part of      |
| 95 / 0.49648               | chloride | 37 / 0.48611            | Small          | 174 / 0.49711              | regulation of cell | 600 / 0.45731 | Disorder of  |
The value of the locality measure of an LP is the same for different coverage thresholds, because the ROCS is the same. However, the coverage threshold influences the number of lexical patterns: the lower a coverage threshold is, the greater the number of LPs are found. For this reason, it is interesting to discuss how new LPs are located in the ontology. In this case, the largest variation happens for the maximum and “3rd Qu.” values, while for the mean and median values this value remains relatively stable. The reduction of the coverage threshold allows for the extraction of patterns with fewer repetitions and closer in the hierarchy, which is a good indicator of the use of naming conventions. For instance, in the DOID the pattern ‘testicular yolk sac tumor’ is found with a mean value of the locality measure of 0.86 for a coverage threshold of 0.1. The size of its ROCS is 10, but it cannot be captured for a coverage threshold 0.5 or 1.0. The analysis of this ROCS reveals that the class ‘testicular yolk sac tumor’, which corresponds to an LP, is the common direct ancestor of other classes in the ROCS like ‘enteric pattern testicular yolk sac tumor’ or ‘macrocystic pattern testicular yolk sac tumor’. However, the lexical pattern could be interpreted as the description of the classes in its ROCS, when the locality is close to 1.0 and there are no classes whose label matches with the LP; in this case, a class with the name of the LP and axioms “subClassOf” linking the new class and the elements of ROCS could enrich the ontology. For this reason, it is worth using a low coverage to obtain many lexical patterns and prioritising them by their locality for obtaining further information about them.

Coverage thresholds have a similar effect on modularity as for locality, because the patterns obtained for lower coverage are more specific and therefore appear in fewer branches defined by the classes of interest. Concerning the distribution of LPs in the contexts defined by the classes of interest, the absolute values of the modularisation measures reveal that, on average, LPs appear in more than one classes of interest. Next, we discuss how they are distributed in such elements using the standard deviation value of the modularisation measure. The high percentage obtained of LPs in G1 is a sign that these ontologies follow the naming guidelines proposed by the OBO Foundry [30]. Except for SNOMED CT, the highest percentage appears in G3, meaning that the lexical content does not appear homogenously distributed in the context defined in the classes of interest, which is a sign of the potential links between the lexical entities of different modules. These results show that modularity permits ontology developers to obtain information about a lexical pattern that could permit them to refine the classes of interest, like including classes in other levels, looking for anomalies or creating links (axioms) between classes.

4.3 Prioritisation of Lexical Patterns

According to our results of the prioritisation of LPs by the standard deviation of the modularisation measure, the GO and SNOMED CT are rich in patterns like ‘disorder of’ or ‘regulation of cell’ that appears in different branches of the hierarchy. These patterns might be inspected to create axioms linking the classes of its ROCS with those that represent the kind of disorder or the cell that is regulated. The Cross-Product Extensions of the GO [15] already include some axioms that are in line with our lexical patterns, which means that our general technique can be useful to carry out a similar process in the GO and SNOMED CT. In the case of the ChEBI and the Human Disease Ontology, further domain knowledge would be needed in order to define axioms from the prioritised LPs.

According to the size of the ROCS shown in Table 2, defining a systematic strategy to assert axioms for the prioritised LPs would affect a high number of classes, increasing the axiomatic richness of the ontologies. Domain experts might be interested in different types of LP rankings according to different criteria or ontology context. The prioritisation of the patterns shown in section 3.5 is just an example of how the metrics can be combined to allow the ontology developer to focus on those patterns with specific properties.

4.4 Future Work

Our future work includes the application of these metrics to all the BioPortal ontologies to extend our lexical characterisation of a bigger corpus and to continue gaining knowledge about the engineering of biomedical ontologies. The combination of all these LP-based metrics could be used as an indicator of the quality of the lexical structure of an ontology, which could be included as additional information in BioPortal or by ontology evaluation frameworks. Besides, our method for extracting LPs can be improved by using more sophisticated natural language processing techniques, and our OntoEnrich software should be able to support the (semi-) automatic creation of axiomatic enrichment scripts using languages like OPPL [31]. Finally, our distance function just takes into account subClassOf axioms, but others, like disjoint axioms [32], or a node-based approach could also be considered. In addition, the analysis of how lexical patterns could improve current semantic similarity measures could be worthwhile.

5. Conclusions

Having more domain content expressed axiomatically should allow reasoning services to process more effectively the content of biomedical ontologies. In this work, we have presented an extension of our lexical analysis method by including locality and modularity metrics. We believe that locality helps to suggest context patterns and modularity helps to focus on specific context of interest for the engineer. These metrics contribute to a better understanding of the engineering of the ontologies and may support domain experts in the prioritising of the most promising parts of the ontologies for axiomatic enrichment.

Acknowledgments

This project has been possible thanks to the Spanish Ministry of Science and Inno-
viation and co-funded by FEDER through grant TIN2010-21388-C02-02 and fellowship BES-2011-046192 (MQM), and to the Fundación Séneca through grant 15295/PI/10.

References

5. Third A. “Hidden semantics”: what can we learn from the names in an ontology? Utica, IL; USA; 2012.
32. Ferreira JD, Hasting J, Couto FM. Exploiting disjointness axioms to improve semantic similarity measures. Bioinformatics 2013 Sep 3; bt3491.