Lexical Characterisation of Bio-ontologies by the Inspection of Regularities in Labels

Manuel Quesada-Martinez¹, Jesualdo Tomás Fernández-Breis¹ and Robert Stevens²
¹ Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, Spain
² School of Computer Science, University of Manchester, UK

Abstract: Hundreds of biomedical ontologies have been produced, with many of the significant, widely used ones being developed in collaborative efforts and following a set of construction principles, which include using a systematic naming convention for their labels. Despite their success, many of these ontologies have lacked a foundation of axioms that would expose the wealth of knowledge in the ontologies to computational reasoning. Our previous results suggest that exploiting the structure of the labels may contribute to an axiomatic enrichment. Hence, in this work we perform a study of the structure of the labels of the ontologies available in BioPortal to classify them in terms of potential interest for axiomatic enrichment.

Keywords: ontology labels, annotations, lexical analysis, OWL, biomedical ontologies, re-use

1. Introduction

The development and use of ontologies is now a mainstream activity within biology and medicine. With the success of the Gene Ontology [1], biomedicine has become one of the most prominent example of the use of ontologies. Currently, ontology search engines like Watson [2] or Swoogle [3] provide access to thousands of ontologies. In terms of repositories, TONES [4] contains more than two hundred, where BioPortal [5], which is the most prominent repository of biomedical ontologies, has a collection of more than three hundred ontologies and controlled vocabularies. It should be noted that a significant proportion of the ontologies available have been developed by domain experts in community efforts [6]. The role of domain experts building ontologies should help the veracity of the domain knowledge, but not necessarily the engineering of the ontology in terms of the axiomatic description of the domain knowledge.

According to [7], an ontology in computer science is a representation of shared background knowledge for a community, which provides the intended meaning of a formal vocabulary (through certain labels). Ontologies should be distinguished from other simple knowledge organization systems because: (1) ontologies define a set of logical axioms with strict semantics; (2) these logical axioms are used to make inferences about the objects being described and to symbolically manipulate knowledge.

Consequently, the axiomatic richness of ontologies determines to what extent its content is machine exploitable, which is critical for their success in applications. For instance, in the case of ontology-based data annotation processes, the existence of logical axioms could provide advanced analysis possibilities since logical axioms provide additional analytical dimensions. However, most current biomedical ontologies have more information within the textual descriptions of the classes than in the form of logical axioms.
In this work, we focus on the automatic identification and analysis of lexical regularities in such textual descriptions. For this purpose, we take advantage of the hidden semantics of ontologies [8], which is the content implicitly codified in the ontology, but not explicitly expressed as logical axioms. For example, the term `oocyte differentiation`, from Gene Ontology, is a type of `cell differentiation` that is implicitly referencing in its label to the term `oocyte` in the Cell Ontology. Creating an equivalency axiom between both classes would permit additional inferences. The Open Biomedical Ontologies Consortium (OBOC) [9] promotes principles as models of good practice (http://www.obofoundry.org/erit.shtml). Three of these principles promote the development of regular, structured labels: (1) consistent formulation of relational assertions in order to prevent inconsistent uses in the definition of the relations (see, for instance, [10]); (2) orthogonality of the ontologies, which promotes the re-use of content existing in other ontologies; and (3) the use of naming conventions for the labels of classes and relations. In our work, these three principles are taken into account in the lexical analysis of ontologies. An example of such systematic naming is that `differentiation` is found in the labels of 807 classes in the Gene Ontology.

In our approach, the lexical analysis of a corpus of ontologies goes through these stages: (1) quantitative and qualitative characterization of the labels; (2) characterisation of the axioms in the ontologies; (3) studying the degree of re-use within our corpus of ontologies, and (4) classifying each ontology in terms of suitability for enrichment.

This paper has therefore two main objectives: (1) extending the framework for the analysis of the structure of the labels in ontologies; and (2) characterisation of the BioPortal ontologies according to the structure of the labels as a validation of (1) above. This will let us know more about the engineering of these ontologies from the lexical perspective, including to what extent the OBOC recommendations are followed. Such a characterisation seeks to identify the suitability of BioPortal ontologies for the application of ontology enrichment methods based on the structure of the labels.

2. Related work

The analysis of the structure of ontologies is challenging. Recently, semi-automatic methods based on detecting regularities have been proposed for ontology quality assurance. These patterns can help to pinpoint anomalies in the structure of the ontology. Using manual methods, domain experts study their ontologies making manual queries to detect irregularities such as missing restrictions [11] [12]. In contrast, frameworks like the Regularity Inspector for Ontologies (RIO) address the automatic detection of these anomalies [13]. RIO detects syntactic and semantic regularities that help users in the identification of terms with shared logical content. In all these cases, the manual, informal detection of lexical regularities guides the search for ontological regularities.

Several recent articles [14] [15] [16] have studied regularities in labels in biomedical ontologies, most focusing on the Gene Ontology (GO). In some approaches, the identification of regularities is considered a step in an ontology enrichment process. For instance, the semantic enrichment of the Molecular Function ontology, which is a part of the Gene Ontology, was addressed in [16]. This process used the manual detection of regularities in labels as a starting point, and extracted the knowledge embedded in the
labels of the classes to carry out the enrichment. Although those results showed that exploiting the hidden semantics within the labels of classes offered significant benefits, the process was tedious and complex, because the manual inspection requires time and a sound knowledge of the domain.

Our recent work has followed some of the guidelines and steps identified in [16]. An initial approach for the automatic detection of lexical regularities in ontologies was described in [17]. The application of the method to five relevant bio-ontologies was partially addressed in [18]. Despite the results confirming the presence of regularities in the labels of those ontologies through automated analysis, our approach (1) lacked the degree of formality required to be able to generalise our studies, and (2) our previous studies did not use a large corpus of ontologies. The work presented in this paper contributes to overcoming these limitations.

The re-use of content from already existing ontologies has traditionally been considered good practice [19]. Despite the existence of tools that support ontology builders and encompass ontology imports, extraction, pruning, refinement, evaluation [20], and proposals for the minimum information to reference an external ontology term [21], this is still a difficult task. Due to the wide variety of available ontologies for supporting the axiomatisation of a particular domain, an ontology builder may have difficulties in deciding which ontology to re-use in a new ontology development. In this work, we will focus on analysing whether we can identify relations between the lexical regularities that can be identified in different ontologies as indicators of possibility for re-use in axiomatic enrichment. The development of methods for checking the logical equivalence between the entities that share the lexical regularities is out of the scope of the present work.

3. Methods

3.1. General issues

Most biomedical ontologies are defined using OWL [1] and OBO [22]. The basic components of an ontology are concepts, properties, individuals and annotations. Both OWL and OBO offer mechanisms to import/re-use ontologies, and both work in the same way. Importing ontologies is transitive, that is, if an ontology \( \theta \) imports an ontology \( \theta' \), the axioms provided by \( \theta' \) will be included in \( \theta \) as well. Consequently, when analysing ontology labels, we will distinguish the main ontology from the imported ones.

**Definition 1** (Main ontology, \( M\theta \)). The main ontology is the one we want to process and analyse, that is, it contains the domain knowledge of interest for the analysis.

**Definition 2** (Imported ontology, \( I\theta \)). The main ontology may re-use content from other ontologies. Such auxiliary ontologies are called imported ontologies and contribute to provide a better conceptualization of the domain and to make

---

1 [http://www.w3.org/TR/owl-ref/](http://www.w3.org/TR/owl-ref/)
the content of the domain ontology more interoperable. The $M\theta$ usually re-uses pieces of each $I\theta$.

Let us illustrate these definitions with the Enzyme Mechanism Ontology\(^2\) (EMO) [23]. EMO describes the domain of the active components of the enzyme’s reaction as well as their roles in reactions. An enzyme is a protein and proteins are composed of amino acid residues. For this reason, the Amino Acids ontology\(^3\), which describes amino acids and their properties, is re-used by EMO. Moreover, the Basic Formal Ontology [24] is also re-used as an upper level ontology.

We include these two definitions because: firstly, it is interesting to study how current biomedical ontologies make use of $I\theta$ as this is an indicator of their degree of re-use; and secondly, we need to contextualize the search of lexical regularities in those cases that the ontology $\theta$ is formed by one $M\theta$ and one or more $I\theta$s. We search lexical regularities just in $M\theta$ because it is the one that should be directly modified by the ontology builders.

### 3.2. Characterisation of axioms

We have previously mentioned that the inspection of regularities could be useful for axiomatically enriching an ontology. In OWL ontologies, representational assertions like classes and properties are represented as logical axioms. The textual descriptions of the ontology entities are usually contained in annotations axiom like label or comment, and sometimes they only appear in the URI fragment of the ontology entity.

Our analysis includes a study of the types of axioms defined in each ontology, since this provides some insights about the engineering of the ontology. Special attention will be paid to annotations, since our method has been designed to analyse the structure of labels. Also, annotations in OWL ontologies are designed to support multilingual definitions for concepts. The internationalisation of labels is another aspect to take into account in the lexical characterisation of ontologies. Obviously, the lexical analysis of ontologies is language-dependent, so a lexical analysis has to be performed for each language used in the ontology.

It should be noted that an axiomatic ontology enrichment method would not aim to create new annotations, but more logical axioms.

### 3.3. Definitions for a lexical characterisation

In [17], a method for taking the labels of the ontology as inputs and creating axioms was proposed. In this work we formalise and expand the idea of lexical regularity, adapting it into the field of ontologies.

Each component of an ontology ($\theta C$) is represented by a Unique Resource Identifier (URI). Ontology concepts are expanded with descriptions in natural language, called annotations, that helps domain experts to better understand their constructs. The content

\(^2\) [http://bioportal.bioontology.org/ontologies/EMO](http://bioportal.bioontology.org/ontologies/EMO)

\(^3\) [http://www.co-ode.org/ontologies/amino-acid](http://www.co-ode.org/ontologies/amino-acid)
of the annotations are designed for human understanding, whereas URIs are for machine processing. Labels are one such type of annotations and are used by ontology builders to express the natural language names of the concepts. Formally, every concept has associated a set of annotations $\mathcal{A}_S = \{A_0 \ldots A_i\}$. Using the Gene Ontology as an example, the concept ’GO_1900365’ has the label ’positive regulation of mRNA’, and ‘GO_1900366’ has the label ’negative regulation of mRNA’.

Fig. 1 Partial representation of the hierarchy of concepts in the Gene Ontology. The graphic shows an example of how the naming of the entities follows certain regularities

Fig. 1 shows a part of the hierarchy of classes and relationships in the Gene Ontology. In this figure the labels are associated with the concepts. It can be seen that the naming of the labels follows a convention that helps humans to understand the meaning of the classes and the relationships. For instance, ’cellular physiological process’ is a child of ’cellular process’ and ’physiological process’. All of them share a part of the label.

3.3.1. Definition of lexical regularities

The formalization of lexical regularities requires the definition of three concepts: the set of delimiter symbols, a tokenise function and tokens.

Definition 3 (Set of delimiters, $\mathcal{D}_S = (l_1 \ldots l_i)$). A set of delimiters is a set of characters expressed in the same alphabet as the labels. These characters permit the decomposition of the labels into smaller fragments.

Definition 4 (Tokenise function, $\mathcal{TF}$). A tokenise function extracts annotations from an ontology entity, and splits them into smaller fragments of text using elements of $\mathcal{D}_S$ as delimiters.
**Definition 5** (Token, $T$). A token is the minimum fragment into which annotations are broken up by a tokenise function.

Once we have formalized these three elements, we can define a lexical regularity in an ontology as:

**Definition 6** (Lexical regularity, $LR$). A lexical regularity is a list of tokens repeated in different entities of $M\theta$. More formally, given an ontology $\theta$ and its $M\theta$ that has the set of components $OCS = \{OC_1 ... OC_i\}$, a lexical regularity is a set of tokens that appear in the same order in annotations associated with different members of $OCS$. The lexical regularity is identified by the sequence of tokens ($LRSOT$) and by the components in which it appears, written $ROCS$, where $ROCS \subseteq OCS$.

To illustrate lexical regularity let us consider the lexical regularity from GO `regulation of` with: $LRSOT = [\text{regulation, of}]$ and the $ROCS = \{GO\_1900365, GO\_1900366\}$. Moreover, the first words of both concepts (`positive` and `negative`) appear in other concepts of the ontology so they are also lexical regularities. Thus, these two concepts can be generalised as `$X Y Z$` where `$X$` would be `positive` or `negative`, `$Y$` would be the action of regulation, and `$Z$` the concept regulated. 1703 labels in GO contain the structure `C binding`, which can be used to extract the binding functions where `C` is a chemical substance. Similarly, 24 types of structural molecule activities have a label with the structure `structural constituent of $M$`, where `$M$` is a macromolecular complex.

### 3.3.2. Properties of lexical regularities

Next, we define two properties of lexical regularities, namely length and frequency.

**Definition 7** (Length of a lexical regularity, $LLR$). The length of an $LR$ is the number of its tokens.

**Definition 8** (Frequency of a lexical regularity, $FLR$). The frequency of an $LR$ is the number of entities in which the regularity appears. Formally, this is the cardinality of the set $ROCS$. $FLR$ can be interpreted as a percentage of ontology entities where it appears.

In the Gene Ontology, `regulation of` is a regularity with $LLR = 2$ and $FLR = 7146$. This regularity has a relative $FLR = 19.55\%$. Moreover, in some cases repetitions of just one word might be of interest, except when they are function words rather than content words. For instance, `binding` is a single word ($LLR = 1$), which encodes some meaning about molecular functions. While large lexical regularities symbolise that the entities where they appear can be easily decomposable into smaller entities, small lexical regularities ($LLR$ close to 1) are also interesting because the entities where they appear should be logically connected.

Another important aspect is whether the lexical regularities correspond to full annotations of ontology entities. For instance, the regularity `binding` is a class of the Gene Ontology. However, other regularities, such as `regulation of`, `positive`,
negative´ or ´involved in´, are not entities despite appearing many times in the ontology.

3.3.3. Filtering lexical regularities

It should be noted that not every group of repeated words embed domain knowledge. Text based approaches usually deploy lists of stop-words to identify words without content meaning and, therefore, are not further processed. Once filtered out, the number of lexical regularities may still be large in some ontologies.

Mechanisms to filter the lexical regularities and highlight the most relevant of them should be useful for ontology developers. In this work, we propose a parameter that takes into account the relative FLR for filtering the results. For example, a regularity with FLR 5 in an ontology with 10 entities is relevant because the relative FLR is 50%. However, in an ontology with 200000 entities it might be irrelevant. This parameter is called the coverage threshold of a lexical regularity.

**Definition 9** (Coverage threshold of a lexical regularity, CTLR): CTLR is the minimum FLR that regularities must reach to be included in the results.

Let us consider that an ontology builder wants to explore the lexical regularities of SNOMED CT, and she considers that regularities repeated in less than 20% of the entities can be filtered out. In this case, the ontology builder would set the parameter CTLR to 20%. In this case, regularities like ´disorder´ (22.3% relative FLR) would be included in the results, but others like ´procedure´ (17.09% relative FLR) would be discarded. Besides, lexical regularities that appear in a small number of labels (e.g., two) are generally filtered, because they are usually insignificant considering the context of a typical biomedical ontology. Concerning the typical size, the BioPortal repository consisted of 368 ontologies with 5,941,254 classes in December 2013.

3.3.4. Hierarchy of lexical regularities

An ontology is built around the hierarchy or taxonomy of concepts it defines. If we consider the OBO principle of systematic naming, this could mean that the label of a concept should contain, to some extent, the label of its parent concept. For instance, if the node ´binding´ defines the general concept of binding, its descendants, which specify types of binding, should contain ´binding´ in their labels. These recommendations drive us to take a special interest in relations between lexical regularities. Apart from the main concept (lexical regularity), two related and derived concepts associated with lexical regularity are also defined.

**Definition 10** (Lexical sub-regularity). A lexical sub-regularity is a sub-sequence of a regularity. The number of sub-regularities for a given regularity is the number of possible combinations of its tokens.

**Definition 11** (Lexical super-regularity). A lexical super-regularity of a given regularity is a regularity that includes the latter one. From a given regularity, super-regularities can be obtained by expanding the regularity in any direction.

---

4 http://www.ihtsdo.org/snomed-ct
The expansion is defined both by the number of tokens and the direction(s) in which the regularity is expanded. A left or right window is the name used to refer to the number of tokens of the expansion to either the left or right.

In the Gene Ontology, the regularity ‘positive regulation of mRNA polyadenylation’ has 15 sub-regularities, including itself. The length of those sub-sub-regularities ranges from 1 to 5. Likewise, ‘positive regulation of mRNA’ and ‘negative regulation of RNA’ are super-regularities of ‘regulation of’, using a left and right window with value 1.

### 3.3.5. Re-use of lexical regularities

Regularities can support the study of how ontologies re-use other ontological resources. In particular, our method is able to suggest potential candidates for re-use by searching for matches for the content of lexical regularities in external ontologies.

Such matches would reinforce the idea that a lexical regularity might be embedding domain knowledge. For instance, the label ‘vitamin binding’ is composed of the tokens ‘vitamin’ and ‘binding’. If the label vitamin appears as the full label of a class of the ChEBI ontology [25], a situation of potential re-use would occur, and this might also contribute to the axiomatic enrichment of our ontology. It should be noted that not only classes could be re-used but also properties. For instance, the lexical regularity ‘encoding’ of the Vaccine Ontology is an object property of the Bone Dysplasia Ontology.

For this purpose we define three types of matches, namely, external exact match, internal exact match and exact regularity match.

**Definition 12** (External exact match, \(EEM\)). An external exact match of a given \(LR\), written \(EEM\), occurs when the label of a class of an external ontology is identical to the \(LR\).

**Definition 13** (Internal exact match, \(IEM\)). An internal exact match of a given \(LR\), written \(IEM\), occurs when an \(LR\) is the label of a class in \(M\theta\).

**Definition 14** (Exact regularity match, \(ERM\)). An exact regularity match of a given \(LR\), written \(ERM\), occurs when an \(LR\) is another \(LR\) in another \(\theta\).

While matches \(EEM\) and \(IEM\) compare \(LRs\) with labels of classes, \(ERM\) compares \(LRs\) with \(LRs\).

### 3.4. Detection of lexical regularities

The method followed to detect lexical regularities in ontologies is depicted in Figure 2. The characterisation of the ontology is influenced by the set of external sources used in the process. Such external sources might be the ontologies available in repositories like BioPortal, which would play the role of semantic context. This semantic context might be composed of ontologies that define a domain close to the domain of the ontology that will be lexically characterized. In general, if the semantic context is not properly

---

[5](http://purl.org/skeletome/bonedysplasia#encoding)
selected, the lexical mappings obtained can be ambiguous due to circumstances such as polysemy. For instance, the search for the term ‘amino acid’ returns 168 hits in 44 BioPortal ontologies, but they might not refer to the same concept.

Fig. 2 The method takes an ontology file as input and detects its lexical regularities by inspecting textual descriptions embedded in the labels of the ontology. The lexical regularities are stored in a database that allows a repository of regularities to be built. We can re-use this repository with other ontologies. The lexical characterization seeks to help domain users to identify repeated content in the whole set of the ontology and in external ontologies.

Our method creates a graph of tokens by processing the labels included in the input ontology. The exploitation of such a graph enables the identification of the lexical regularities and the links between them. Each node of the graph corresponds to one token. In addition, each node may be linked to other nodes through an arrow. Such arrows symbolise that nodes appear consecutively and in that order in a label.

In our method, finding a lexical regularity of length \( N \) requires navigating through \( N - 1 \) edges starting from an initial node. We can obtain the whole set of lexical regularities within an ontology by repeating the process in all the nodes of the graph. Fig. 3 shows three labels using our representation. The arrows represent edges and the numbers refer to the identification of the labels where in it appears. We can see that the regularity ‘regulation of’ (length 2) can be found. The node regulation has input arrows from labels 1 and 2. This shows that it appears in those labels. The labels 1 and 2 of the arrow between ‘of’ and ‘isoprenoid’ represents that both words appear consecutively in these two labels. ‘Of isoprenoid’ is a lexical regularity of length two and with two repetitions. Consequently, ‘regulation of isoprenoid’ is a lexical regularity that appears in labels 1, 2 but not in 3. The graph also includes information about the position in the label (e.g., the same word could appear several times in the same label), together with the URI of the class, etc.
Fig. 3 Example of the graph built using a subset of four labels extracted from GO. Underlined boxes represent initial nodes; solid white boxes stand for the initial node of labels; and solid grey boxes are neither initial nor final nodes. A node can play different roles at the same time (e.g. “isoprenoid” is both an initial and intermediate node).

4. Result

We used ontologies from the BioPortal [5] repository in our case study because it is the largest repository of biomedical ontologies. The ontologies included in this study are the ones publicly accessible in OWL format in November 2012 (178 ontologies from which 19 were discarded due to their importing inaccessible OWL files and another 41 were discarded due to the absence of labels on their entities). In this section, we summarise the main results of our experiment. The complete results can be found at http://miuras.inf.um.es/biotest. The ontologies were analysed by using the OntoEnrich software [17], which is able to process ontologies in OWL.

The results include: the types of axioms in the ontologies; the structure of the labels; the properties of the lexical regularities identified in those labels; and our classification of BioPortal ontologies in terms of their suitability for applying axiomatic enrichment methods.

4.1. Imported ontologies

In this subsection we study how ontologies make use of others in the definition of domain concepts. We have analysed the number of ontologies imported by each ontology in our corpus. For this purpose, the functions provided by the OWL API [26] are used.

Each ontology imports 1.77 ontologies on average. The ontology with the largest number of imports (221 imported ontologies) is the Semantic Web for Earth and Environment Technology [27], which is a middle-level ontology created using the
principle of data interoperability by supporting knowledge re-use. It is focused on applications for the discovery and use of Earth science data.

The group of ontologies that makes imports contains 23% of the ontologies. 90% of the ontologies of this set import more than one ontology. The three most frequently imported ontologies (Θs) are the ‘Basic Formal Ontology’, the ‘OBO Relation Ontology’ and ‘The Information Artefact Ontology’. Generally, the ontologies imported by BioPortal ontologies are structural ones. Structural ontologies help ontology builders in development of content by providing general knowledge, such as time metrics, space metrics or formal components that help in the specification of the domain – upper level ontologies such as BFO and associated relationships.

### 4.2. Characterization of axioms

We have analysed which types of axioms are the most frequent in the BioPortal ontologies. Our results reveal that ‘AnnotationAssertion’ is the most frequent type of axiom (57.63%), followed by ‘SubClassOf’ (26.84%). These two types of axiom represent more than 75% of 35 types.

<table>
<thead>
<tr>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0003</td>
<td>0.98</td>
<td>0.93</td>
<td>3.91</td>
</tr>
</tbody>
</table>

Table 1 Summary of the ratio of the number of labels / number of classes, considering the ontologies of our corpus. This ratio measures the richness and distribution of labels. The mean and median values show that the majority of the ontologies are rich in labels having one label annotation for each class.

We have characterized the use of annotations, finding that 41 ontologies do not have annotations. For those ontologies with annotations, we have studied the number and distribution of labels; an ontology class could have more than one associated label. For this purpose, we have calculated the ratio ‘number of labels / number of classes’. Table 1 shows the summary of the results. We inspected which languages are used in our corpus. Table 2 shows the results obtained from our corpus.

The distribution of those ontologies with the ratio under or over 1 is shown in Fig. 4. A class has more than one label when: (1) synonyms are provided, so these labels are defined in the same language; (2) labels in different languages are provided. Three ontologies have more than one label per class, namely, ‘PHArmacogenomic Relationships Ontology’ (3.92), ‘Skin Physiology Ontology’ (1.98), and ‘Epilepsy’ (1.87). The ‘PHArmaco genomic Relationships Ontology’ and the ‘Skin Physiology Ontology’ both have their labels expressed in the same language, so they might be examples of case (1). On the contrary, the ‘Epilepsy ontology’ is an example of case (2) as it has more than one label per class and the labels are codified in English (‘en’) and Portuguese (‘pt’).

<table>
<thead>
<tr>
<th>Language</th>
<th>Percentage out of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undefined</td>
<td>75.63% (961276)</td>
</tr>
<tr>
<td>‘en’</td>
<td>24.35% (309448)</td>
</tr>
<tr>
<td>‘en-us’</td>
<td>0.01% (157)</td>
</tr>
<tr>
<td>‘pt’</td>
<td>0.01% (123)</td>
</tr>
</tbody>
</table>

Table 2 Languages used in the BioPortal ontologies and its distribution of labels by language. Mainly, 75% of the annotations do not specify the language, and we assume that they are in the same language of the ontology.
Fig. 4 Ontologies of our corpus grouped according to the value of the ratio number of labels / number of classes.

In contrast, the ‘Eagle-I Ontology’ has less than one label per class and has annotations in three different languages: ‘en’, ‘en-us’ and ‘undefined’. In these cases, the meta-information about the language could be unified.

4.3. Characterisation of lexical regularities

We have applied our method to each ontology in our corpus. We have analysed lexical regularities with different values of $CTL_R$: from 0.2% to 1.0% with increments of 0.2 and from 1.0% to 5.0% with increments of 1.0. Next, we show the results with the coverage set to 1%.

For each ontology, its lexical regularities were extracted and analysed. Then, the results were analysed to classify the ontologies into groups of interest for enrichment by applying agglomerative hierarchical and k-means clustering techniques. The variables used are:

- **Number of labels**: number of labels in an ontology.
- **Number of lexical regularities**: number of lexical regularities found in an ontology.
- **Classes affected by lexical regularities**: this is considered in both absolute and relative terms. For this reason this variable is divided in two sub-variables that stand for: (1) the number and (2) the percentage of classes in which the lexical regularities are found.
- **Classes affected by matches**: number of classes where LPs with exact matches are found; it is considered in both absolute and relative terms.
- **Repetition of words**: for each ontology, measured how many different words exist in the labels.
- **Coverage threshold of a LR ($CTL_R$)**.

4.3.1. Description of lexical regularities

We have calculated that the mean value of repeated words in labels is 81.67%, ranging from 50.01% to 94.7%. This can be interpreted as a sign of regularity, and the study of these regularities may be useful for the axiomatic enrichment of the ontology classes. Table 3 shows the global descriptors for the dataset composed of 118 BioPortal ontologies (introduced at the beginning of the Section 4). The first row is about the
lexical regularities. We show the total number of lexical regularities and how many are unique. We found that 23.49% of the regularities appear in more than one ontology. This suggests that the axiomatisation of these labels is a potential point of re-use between ontologies (see \textit{ERM}).

<table>
<thead>
<tr>
<th>Number of LRs</th>
<th>Total</th>
<th>8175</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unique</td>
<td>6254</td>
</tr>
<tr>
<td>Length (LLR)</td>
<td>Max</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>X50</td>
<td>1</td>
</tr>
<tr>
<td>Frequency (FLR)</td>
<td>Max</td>
<td>47623</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>83.12</td>
</tr>
<tr>
<td></td>
<td>X50</td>
<td>4</td>
</tr>
<tr>
<td>Matches</td>
<td>\textit{EEM}</td>
<td>36.44%</td>
</tr>
<tr>
<td></td>
<td>\textit{IEM}</td>
<td>15.60%</td>
</tr>
<tr>
<td></td>
<td>\textit{ERM}</td>
<td>23.49%</td>
</tr>
</tbody>
</table>

\textbf{Table 3} Numerical metrics about the features of the lexical regularities. Rows 2 and 3 show the maximum, mean and median values for the metrics LLR and FLR. Row 4 shows the percentage of lexical regularities where matches were found.

Fig. 5 shows the percentage of lexical regularities (out of the total number of labels in the ontology) and the percentage of classes covered for each ontology analysed. In general, both variables have a similar trend, except for those cases where the lexical regularities exceed the number of labels. Statistically speaking, these variables show a positive correlation ($r=0.938$, $p=0.000$) using the Pearson's $r$ test.

\textbf{Fig. 5} Graphical representation of (1) percentage of LRs (out of the total number of labels in the ontology) and (2) percentage of classes whose label has lexical regularities. Ontologies are shorted according to the clusters. Vertical dotted lines delimiter each cluster and they are shown in the order cluster1, cluster2 and cluster3.
Fig. 6 shows the number of LPs with one or more EEM and their relation with the total number of LPs for each ontology. As expected, there is also a positive correlation ($r=0.765$, $p=0.000$). A positive correlation between the number of lexical regularities and the repetition of words is significant ($r=0.41$, $p=0.000$). However, the correlation between the number of classes affected by the regularities and the repetition of words is found negative but not significant ($r=-0.118$, $p=0.246$).

![Graphical representation](Image)

**Fig. 6** Graphical representation of (1) number of lexical regularities and (2) number of those lexical regularities with external matches (EEM). Ontologies are sorted according to the clusters. Vertical dotted lines delimiter each cluster and they are shown in the order cluster1, cluster2 and cluster3.

### 4.3.2. Clustering

We have clustered the set of ontologies by the regularities. We performed first an agglomerative hierarchical clustering. The inspection of the dendrogram suggested the existence of three differentiated groups of ontologies. We then selected the clustering variables: percentage of classes for which regularities 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 these variables, obtaining three groups, whose centroids have the values shown in Table 4.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cluster1</th>
<th>Cluster2</th>
<th>Cluster3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularities</td>
<td>81.73</td>
<td>20.85</td>
<td>50.03</td>
</tr>
<tr>
<td>Matches</td>
<td>19.16</td>
<td>2.63</td>
<td>9.90</td>
</tr>
<tr>
<td>Repetitions</td>
<td>72.96</td>
<td>62.85</td>
<td>73.45</td>
</tr>
<tr>
<td>Ontologies</td>
<td>42</td>
<td>24</td>
<td>33</td>
</tr>
</tbody>
</table>

**Table 4** Description of the clusters
Cluster1 includes ontologies whose percentage of classes with regularities is high and those with the highest percentage of classes that contain LRs with matches. The ontologies in this cluster are the most suitable for applying enrichment methods because the LR can help to systematically enrich the classes where it appears, and if the LR is matched then it could be re-used in the enrichment. Cluster2 includes ontologies with low scores for both regularities and matches and, therefore, it includes the least suitable ontologies for the enrichment process (the low lexical regularity and few matches mean the knowledge implicit in the text is difficult to extract by our means). Finally, Cluster3 includes those ontologies with an intermediate score for class labels formed from regularities.

The 99 ontologies for which regularities were found are distributed as shown in Table 4: 42 in Cluster1, 24 in Cluster2 and 33 in Cluster3. The members of each cluster can be seen on the website http://miuras.inf.um.es/biotest.

5. Discussion and conclusions

In this paper we have presented an approach for the characterisation of ontologies by the structure of their labels. Previous work in the biomedical ontology community revealed that biomedical ontologies contain implicit knowledge in the text of their labels and, therefore, have semantics hidden from computational support. Thus the hypothesis is that an automatic detection and characterisation of such lexical regularities can be made and afford information for axiomatic enrichment. Our method depends on regularity or systematicity in labels, so the enrichment will not work if labels are not regular. The results of the application of the method to BioPortal ontologies confirm this hypothesis, and we think that an automatic enrichment approach could work due to the following major findings.

The majority of the BioPortal ontologies do not re-use concepts from other ontologies through imports, so detecting common lexical regularities could potentially re-usable entities. 77.22% of the ontologies do not import any other ontology, and the most imported ontologies are structural ontologies like BFO that support the definition of the ontology but add no new concepts related to the domain. Although it is outside the scope of the current study, it would be interesting to analyse whether ontologies that import BFO show more lexical regularities and if such regularities include the terms of BFO. Our approach measures the re-use in terms of explicit imports, which could be complemented with the study of use of URIs from non-explicitly imported ontologies.

The fact that most matches have been found in external ontologies suggests that the re-use of this content for enriching current biomedical ontologies may be a significant contribution. 15.60% of the lexical regularities correspond to full labels of classes in the ontologies; that is, class labels are compound terms including labels from other classes. In addition, 36.44% of these labels correspond to classes from external ontologies (see Table 3). These results suggest that the number of links between ontologies is lower than the degree of potential relation between the content of different ontologies. The existence of these full matches with entities from external ontologies means that there is potentially re-usable content, which would improve the orthogonality of this corpus of biomedical ontologies. It should be noted that matches must be considered as suggestions, since no logical equivalence is computed. However, this kind of lexical
analysis, coupled with inspection by an ontology author, has the potential to facilitate
the axiomatic enrichment of ontologies, exposing more of the knowledge implicit in
labels to computational processing.

75% of the axioms of BioPortal ontologies are annotations or SubClassOf, which we
interpret as showing that most BioPortal ontologies are taxonomies structures and are
rich in natural language content. We think that these ontologies should have more
axioms of other types, which would enrich the ontologies by providing a more precise
computational formalisation of the meaning of the concepts, and such additional content
would be machine processable. We think that the analysis of the structure of the labels
is not enough for this task, although enrichment processes should also benefit from the
fact that there might be lexical regularities that are names/labels of classes and
properties in other bio-ontologies. For instance, we think that if the label of a class
contains a lexical regularity, which matches the full label of another class, this indicates
that both classes might be related. Consequently, an axiom linking both classes might
be defined. Unfortunately, our current method is so far not able to determine whether
the matches are semantically correct, so human intervention would be required.

An ontology is potentially suitable for enrichment if its labels contain regularities, such
regularities affect many classes in the ontology, and the regularities have matches with
content from external ontologies. The cluster analysis reveals that Cluster1 and Cluster3
contain ontologies with such properties, so 75% of the BioPortal ontologies analysed
could benefit from enrichment processes.

We find that 41 ontologies do not define labels for the concepts, which suggests that the
developers of such ontologies are defining such information in the URI of the ontology
entities, which is a bad practice, since the URI is not a content designed for human
consumption. The analysis of the ontologies with labels shows that the ratio of
labels/classes is close to 1, which means that the ontology builders are providing a
textual definition of each entity. This also means that defining synonyms is not a
common practice. An indicator of such a situation would be to have a higher ratio of
labels/classes. Having multiple labels per class might also mean that the ontology has
been internationalised, that is, labels in different languages are defined. However,
according to our result this is not a frequent practice in this corpus of ontologies, and
most ontologies have no language associated with the labels.

As previously mentioned in this paper, biomedical ontologies have traditionally been
mainly used for supporting annotation processes. Such annotations facilitate the
understanding and sharing of the meaning of the data, regardless of the language in
which the label is written. Hence, providing labels in multiple languages would
facilitate such understanding for international communities of practice. It is obvious that
English is the language of science, but other languages are necessary for medicine. For
example, SNOMED CT, which is the most important clinical terminology, has
developed translations to several languages to facilitate its use while sharing the clinical
concepts.

For the detection of the whole set of lexical regularities, we considered options like n-
gram or high performance graph databases. However, 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. The structure of the graph allows
pruning of the search space during the detection of the lexical regularities, thus producing faster answers for queries. For instance, the coverage threshold allows pruning of the search when the algorithm explores a branch that does not reach the minimum coverage value. We also considered using other approaches based on Natural Language Processing (NLP), but annotations are accurate definitions, usually composed of single sentences. However, we think that using POS-tagging techniques to include the type of words in lexical characterisation would allow a more sophisticated and precise analysis.

Our label analysis method represents hierarchies of lexical regularities through sub or super relations, which generate complexity; not only for the detection and representation of such relations, but also in the navigation. On the one hand, the time required in the detection and navigation through the hierarchy of lexical regularities increases with the number of labels of the ontology. For this reason, the use of an accurate organisation of the tokens in a label is required in order to process ontologies with thousands of labels and words. On the other hand, the navigation should permit users to access super regularities as quickly as possible or vice-versa, and this has to be provided by the algorithm that searches lexical regularities. Obviously, sub-regularities appear in more labels than their related super-regularities, and this aspect has to be taken into account by the method.

Although the time required in the extraction of the regularities has not been included as a variable of study in this work, it should be noted that we use a predefined timeout for the analysis of the ontologies. Hence, if the time exceeds such a timeout the analysis stops. This permits us to analyse the effect of using different parameters such as the coverage threshold. This has permitted us to see that the analysis of SNOMED CT exceeds the timeout with very small coverage thresholds because of its huge number of labels and because such a low threshold would retrieve lexical regularities with few repetitions. This is why defining an appropriate coverage threshold has an impact on both the time of analysis and the relevance of the lexical regularities identified.

Lexical regularities in the labels could be used to identify potential modules in the ontology. This solution would address the analysis in a part of the ontology that shares lexical content, regardless of the size of the ontology. This would permit execution focused analyses on the ontologies, which would speed-up the analysis of large ontologies like SNOMED CT.

We think that the analysis of the relations sub-regularities/super-regularities together with the taxonomic links asserted in the ontology should provide interesting information about the engineering of the ontology, since new taxonomic links could be suggested (as has been traditionally done for different purposes in ontology engineering), and to measure the degree of use of a systematic naming convention.

We approach axiomatic enrichment processes as described in [16]. There, the detection of lexical regularities was one phase of the enrichment process. The next step would be to design knowledge patterns by using the information of the lexical regularities and the matches to other concepts. The final implementation and execution of the knowledge patterns as OWL axioms could be done using scripting languages like OPPL [28].
In summary, we have been able to provide some formalisation of our framework for analysing the structure of the labels of ontologies. This framework has been applied to the most popular corpus of biomedical ontologies, BioPortal. Our results provide interesting information about the engineering of these ontologies and to what extent best practices and principles recommended by the OBO Foundry are followed. We see this kind of approach as another piece in the software landscape for advanced ontology engineering. Our next goal will be to use the information provided by this lexical analysis to define a practical method for ontology enrichment.

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 and the Fundación Séneca through grant 15295/PI/10. The Spanish Ministry of Science and Innovation funds Manuel Quesada-Martínez through fellowship BES-2011-046192.

References


