Studying the Reuse of Content in Biomedical Ontologies: An Axiom-Based Approach

Manuel Quesada-Martínez and Jesualdo Tomás Fernández-Breis(✉)

Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca,
30100 Murcia, Spain
{manuel.quesada,jfernand}@um.es

Abstract. The biomedical community has developed many ontologies in the last years, which may follow a set of community accepted principles for ontology development such as the ones proposed by the OBO Foundry. One of such principles is the orthogonality of biomedical ontologies, which should be based on the reuse of existing content. Previous works have studied how ontology matching techniques help to increase the number of terms reused. In this paper we investigate to what extent the reuse of terms also mean reuse of logical axioms. For this purpose, our method identifies two different ways of reusing terms, reuse of URIs (implicit reuse) and reuse of concepts (explicit reuse). The method is also able of detecting hidden axioms, that is, axioms associated with a reused term but that are not actually reused. We have developed and applied our method to a corpus of 144 OBO Foundry ontologies. The results show that 75 ontologies implicitly reuse terms, 50% of which also explicitly does it. The characterisation based on reuse enables the visualisation of the corpus as a dependency graph that can be clustered for grouping ontologies by their reuse profile. Finally, the application of a locality-based module extractor reveals that roughly 2 000 terms and 20 000 hidden axioms, on average, could be automatically reused.

Keywords: Biomedical ontologies · Ontology axiomatisation · Reuse

1 Introduction

The biomedical community has now developed a significant number of ontologies. The curation of biomedical ontologies is a complex task and they evolve rapidly, so new versions are regularly and frequently published in ontology repositories. Ontologies should play a critical role in the achievement of semantic interoperability in healthcare, as it was stated by the Semantic Health Net[1]. Therefore, the *quality assurance* of the content of biomedical ontologies is important, but it is becoming harder and harder due to the increasing number and size of biomedical ontologies. Briefly speaking, ontologies describe a domain using terms/classes, properties and instances that are implemented using a formal language.

[1] http://www.semantichealthnet.eu/.

© Springer International Publishing AG 2017
A. ten Teije et al. (Eds.): AIME 2017, LNAI 10259, pp. 3–13, 2017.
DOI: 10.1007/978-3-319-59758-4_1

Ontology entities have natural language annotations that make them understandable by humans, but such meaning is provided to the machines in the form of logical axioms.

The OBO Foundry [10] promotes as a set of principles for building ontologies. One of these principles promotes the reuse of terms for building an orthogonal set of ontologies[2]. Orthogonality could be used when terms can be jointly applied to describe complementary but distinguishable perspectives on the same biological or medical entity. The reuse in biomedical ontologies has been studied in works like [3,7–9]. In [9] the analysis of prominent case studies on ontology reuse was performed, discussing the need for methodologies that optimally exploit human and computational content when terms are reused. Later, in [3] the level of explicit term reuse among the OBO foundry ontologies was studied. Recently, a systematic analysis of term reuse and overlap has been performed in (1) Gene Ontology and (2) between other biomedical ontologies [7,8]. However, those works mainly focused on analysing and promoting term reuse but did not analyse the reuse of axioms. In general, the more axioms the ontology has, the more inferencing capability it has. Hence, the goal of this work is to provide insights in how the reuse of logical axioms can be improved.

2 Methods

2.1 Types of Term Reuse in Biomedical Ontologies

The reuse of content is a best practice included in methodologies for building ontologies [9] and it is one of the principles proposed by the OBO Foundry. As mentioned, orthogonality permits ontology developers to focus on the creation of the content specific of a given subdomain, and to include content from other subdomains by reusing properties or axioms. According to the OBO Foundry principle, ontology terms can be reused in different ways:

- **Explicit reuse of full ontologies:** options for importing ontologies of languages such as OWL permits to have access to their entities and axioms[3]. The `owl:imports` operation is transitive, which means that if an ontology θ_1 imports the ontology θ_2, and θ_2 imports θ_3, then θ_1 imports the content of θ_2 and θ_3. The *import closure* of an ontology θ is the smallest set containing the axioms of θ and all the axioms of the ontologies imported by θ [2]. For an ontology θ we define two sets of classes θC and θC_{IC} where θC contains all the classes directly defined by θ and θC_{IC} the classes imported from external ontologies. We consider that a term is explicitly reused when it comes from an imported ontology.
- **Implicit reuse of individual terms:** this can be done by reusing the term URI (Uniform Resource Identifier) without importing the ontology.

The reuse of ontology content requires a *source ontology* and an *external* one. Figure 1 shows the axiomatic definition of the term *Cleavage: 16-cell*[4]. This

[2] http://www.obofoundry.org/principles/fp-001-open.html.
[3] https://www.w3.org/TR/owl2-syntax/#Imports.
[4] http://purl.obolibrary.org/obo/ZFS_0000005.

term is originally defined in the Zebrafish Developmental Stages Ontology (ZFS) (Fig. 1 right) and it is implicitly reused in the Zebrafish Anatomy and Development Ontology (ZFA) (Fig. 1 left). In this example, ZFA plays the role of *source ontology* and ZFS is the *external* one. In this example only the URI is reused, since the axioms defined in ZFS are not available in ZFA. Thus, the implicit reuse of `ZFS_0000005` does not imply reusing the axioms: `part of some cleavage` or `immediately_preceded_by some Cleavage:8-cell`. This means that a tool using ZFA could not use these two axioms to make inferences. In this work, we will refer to these axioms as *hidden axioms*.

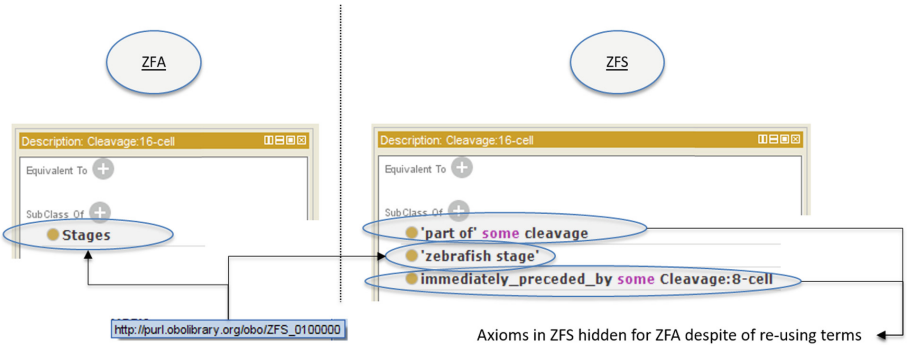

Fig. 1. Axiomatic definition of the term *Cleavage: 16-cell* (ZFS_0000005). (Right) Axioms associated with the term in the original ontology (ZFS). (Left) Axioms associated with the term in the ZFA ontology, which implicitly imports the term through its URI.

2.2 Characterisation of Ontologies Based on Reuse

Ontologies can be characterised according to the type of reuse they exhibit. The relation between a *source ontology* and *external ontologies* is usually 1:m. Figure 2 shows three examples of the behaviour followed by three ontologies extracted from the OBO Foundry repository: ZFA, the Comparative Data Analysis Ontology (CDAO) and the Cephalopod Ontology (CEPH). Dark circles represent the *source ontologies*, and the number of terms with local URI are shown in brackets. White circles represent the *external ontologies*, dotted circle lines mean implicit reuse and solid circle lines mean explicit reuse. For example, CEPH defines 325 terms, and it reuses terms from the Uberon Multi-Species Anatomy Ontology (UBERON): 72 implicitly and 408 explicitly reused.

Therefore, an ontology can be classified in one of the following groups: (1) no reuse, (2) implicit reuse, (3) explicit reuse, and (4) implicit and explicit reuse. In the running example, ZFA, CDAO and CEPH belong to groups 2, 3 and 4 respectively. The explicit importation of one ontology does not necessarily imply that the content of one ontology is reused. This does not mean to reuse the whole

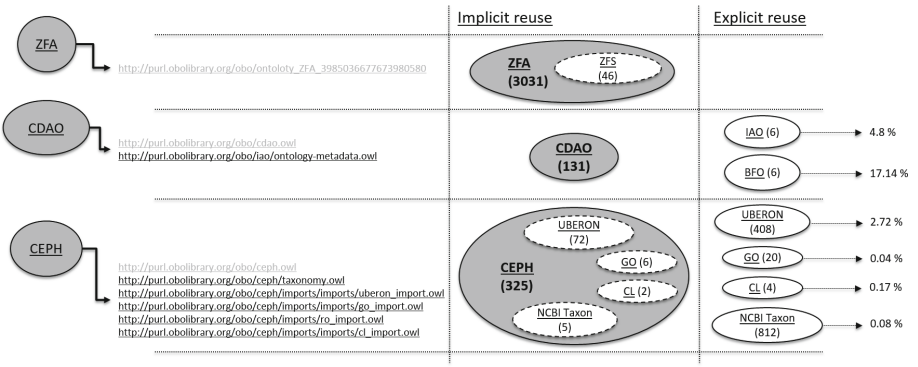

Fig. 2. Example of the method of reuse between of three ontologies in the OBO Foundry repository.

content of the original ontology either, as the import file could just include a fragment created with the purpose of being reused. For example, in Fig. 2, CEPH explicitly reuses less than 3% of the terms defined in the *external ontologies*.

2.3 Identification of Hidden Axioms

We propose a method to measure how much of the potentially reusable logical knowledge is actually reused. For this, we follow the next steps:

1. **Analysis of content driven by URIs:** Analysis of the identifiers of the *source ontology* entities, assuming that the OBO Foundry principle of URI / Identifier Space[5] is followed. This principle defines that the URI of each term is the concatenation of the ontology base URI (prefix) and an identifier. For example, *Cleveage: 16-cell* in Fig. 1 has the prefix `ZFS` and the identifier `0000005`. We process term URIs by applying a regular expression[6]. The analysis groups terms in *reused sets*, which are groups of terms defined in a *source ontology* or in its *import closure* and that share the prefix (white circles in Fig. 2).
2. **Retrieval of the *external ontologies*:** The method needs the complete ontologies that are reused in order to calculate how much content is actually reused. For example, if ZFA implicitly reuses ZFS, then the method needs to process the complete ZFS ontology.
3. **Creating axioms sets:** For each *reused set* we create two sets of axioms, one for the axioms included in the *source ontology*, and another one for the axioms included in the *complete external ontology*.
4. **Finding hidden axioms:** For each *reused set*, the axioms of the *complete external ontology* that are not included in the *source ontology* are considered *hidden axioms*.

[5] http://www.obofoundry.org/principles/fp-003-uris.html.
[6] ^http://purl.obolibrary.org/obo/([A-Za-z]+)_(\\d+).

2.4 A Modular Strategy for Increasing the Amount of Knowledge that is Already Being Reused

Finally, we want to propose an automatic mechanism that exploits the information provided by our method to increase the amount of knowledge that is already being reused. We propose the use of mechanisms for the automatic extraction of ontology modules [4,6]. In particular, we propose to use locality-based modules[7]. A locality-based module M is a subset of the axioms in an ontology θ, and is extracted from θ for a set S of terms (class or property names). The set S is called a *seed signature* of M. Informally, everything the ontology θ knows about the topic consisting of the terms in S and M is already known by its module M. The remainder of O knows nothing non-trivial about this topic.

We propose to extract modules of the *complete external ontologies* using as *seed signature* the classes reused by the *source ontology*. This will axiomatically enrich the *source ontology* using the minimum amount of logical content linked to the reused terms. The module could include new axioms but also new terms. For example, if the axiom `part of` of Fig. 1 right is reused, then the term *cleavage* from ZFS would be included too.

3 Results

3.1 Experimental Setup

We analysed the OBO Foundry ontologies publicly available at[8]. The corpus was formed by 144 ontologies. For each ontology, we processed the latest version available in BioPortal [11]. In case such ontology was not available in BioPortal we tried to download the file through the PURL[9] address. The ontologies were downloaded in January 2017. Our automatic process was not able to obtain 3 out of the 144 ontologies in OWL format. We used the OWL API [5] for the manipulation of the ontologies. The method was implemented in Java by using a shared memory algorithm. The method was executed using 64 processors and 300 GB RAM. The processing time was 2.5 h (download time not included). 18 out of 141 ontologies could not been loaded by the OWL API due to inaccessible import references or unparseable content. As a result, we analysed 123 ontologies.

Next, the major results are described. The complete description of the corpus and further results can be found at our website[10].

3.2 Analysis of the Reused Terms URIs

63 ontologies correctly applied the OBO principle explained in Sect. 2.3 to define their URIs. 60 ontologies contained terms that do not follow the principle: 55

[7] http://owl.cs.manchester.ac.uk/research/modularity/.
[8] http://www.obofoundry.org/.
[9] https://github.com/OBOFoundry/purl.obolibrary.org/.
[10] http://sele.inf.um.es/ontoenrich/projects/reuse/aime2017/.

ontologies had such cases only for implicitly-reused terms, 5 for only explicitly-reused ones, and 5 ontologies had cases for both types of reuse. Table 1 shows 5 examples of such situations.

Table 1. Example of URIs that do not follow the format proposed by the URIs/Identifiers principle of the OBO Foundry.

	Source ontology	Incorrect URI
E.g. 1	CDAO	http://www.geneontology.org/formats/oboInOwl#DbXref
E.g. 2	Chemical Inf. Ont.	http://semanticscience.org/resource/CHEMINF_000318
E.g. 3	Chemical Inf. Ont.	http://www.ifomis.org/bfo/1.1/snap#GenericallyDependentContinuant
E.g. 4	Cell Line Ont.	http://www.ebi.ac.uk/cellline#cervical_carcinoma_cell_line
E.g. 5	Cell Line Ont.	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Iliac_Vein

3.3 Analysis by the Type of Reuse

Figure 3 shows the distribution of ontologies by the type of reuse. 49 ontologies did not reuse terms. The remaining 75 ontologies imported at least one term, with implicit reuse being the prominent strategy. The explicit reuse of terms is commonly combined with the implicit one, so ontology developers integrate the reused terms in the *source ontology*, and perform some enrichment with external content.

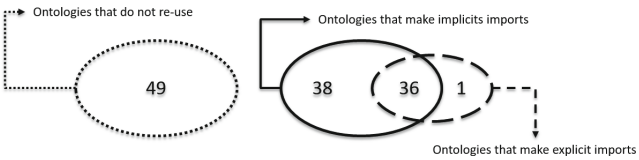

Fig. 3. Distribution of the ontologies according to the type of reuse that they perform.

We used the data obtained by our method as input of the The Open Graph Viz Platform[11]. We built a graph, which can be visualised and explored from different focuses by using filters, for example: (1) Fig. 4 left highlights those ontologies that reuse OBI and other ontologies and relations can be observed in the background; (2) Fig. 4 right shows the filtered graph based on a clustering algorithm that is explained next.

It should be pointed out that the nodes represent ontologies. The directed edges between two nodes means that the node from which the edge departs is the *source ontology* and the other is the *external* one. The weight of each edge represents the number of terms reused, which is represented by the thickness of the edges. The size of each node represents the number of times that the ontology

[11] https://gephi.org/.

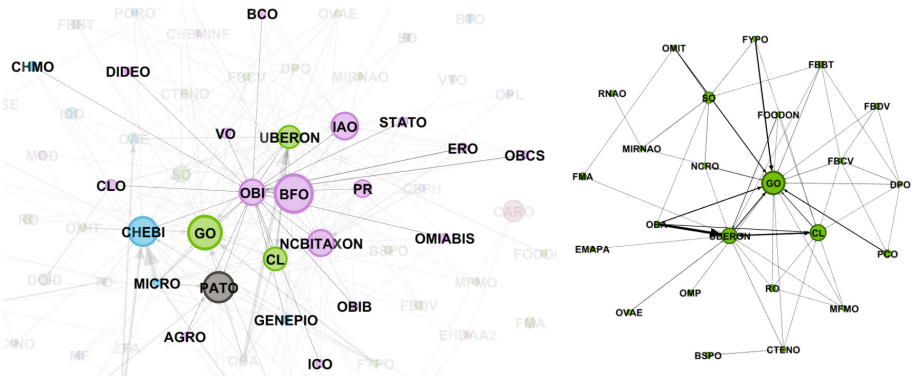

Fig. 4. Graphs that represent the reuse between the ontologies in our corpus. Generated with Gephi using Fruchterman Reingold as layout algorithm to minimise overlap.

is reused. Using Fig. 4 right as example, the Gene Ontology (GO) is reused by the Ontology of Biological Attributes (OBA), GO is reused more times than OBA, and OBA reused more terms from UBERON than from GO. Finally, we performed a cluster analysis of the ontologies using as parameter the weight of the edges (see report at[12]). Clusters are represented by colours in the graph.

The cluster analysis returned 51 clusters. More than 60% of the ontologies were classified in 9 clusters; the reminder clusters had just one member, what means that they do not reuse content. Conceptually, the clusters can be used, for example, to visualise: (1) groups of ontologies that reuse a similar number of terms between them (Fig. 4 right), (2) groups of ontologies that are frequently reused by others, or (3) a small set of ontologies with a high reuse between them in comparison with the members of other clusters (see more figures with the clusters in our webpage). Visualisations like these might contribute to the understanding of the reuse among a large set of ontologies, and they offer different perspectives of analysis to ontology developers.

3.4 Analysis of Hidden Axioms and Terms Already Reused

Finally, we analysed the existence of *hidden axioms* associated with relations already reused in our corpus. Figure 5 summarizes to what extent the reuse of axioms is performed and how the application of the modularity algorithm could be used to increase the reuse of terms and axioms. This result comes from analysing both the implicit and explicit reuse.

– *Terms reuse:* Fig. 5 left compares the mean number of terms that are reused and the potentially reusable ones from the *external ontologies*. The mean number of terms implicitly reused by the analysed ontologies is 855 and the number of explicitly reused ones is 1 210 terms. This difference makes us think

[12] http://sele.inf.um.es/ontoenrich/projects/reuse/aime2017/cluster.

that the `owl:import` operation is not including all the content from the original ontology but a simplified version (e.g. see the percentage of the explicit reuse shown in Fig. 2). The application of our modular strategy finds that the signature of the automatically obtained modules, which were extracted using as *seed signature* already reused terms, contains a mean of 2 016 and 2 376 terms respectively for implicit and explicit reuse. The modules can be imported containing terms logically link to the one reused.

– *Axioms reuse:* Fig. 5 right performs a similar analysis, but focused on axioms instead of terms. The mean number of axioms associated with implicitly reused terms in *source ontologies* (also existing in the external ones) is 2 390, whereas the mean number of axioms associated with such terms only in the *external ontologies* is 22 680; this means that, on average, each ontology has 20 290 *hidden axioms*. The results for explicitly reused terms is, respectively, 2 690 reused axioms and 27 710 hidden ones.

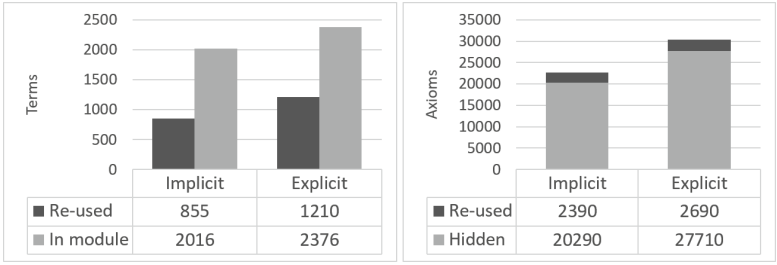

Fig. 5. (Left) Comparison between the number of reused terms and those included in the locality-module extracted. (Right) Comparison between those axioms reused and *hidden axioms* in the complete *external ontology*.

Finally, Fig. 6 shows the most frequent axioms linked with terms reused in the *source ontologies* (left), and that are hidden in the *external* ones (right).

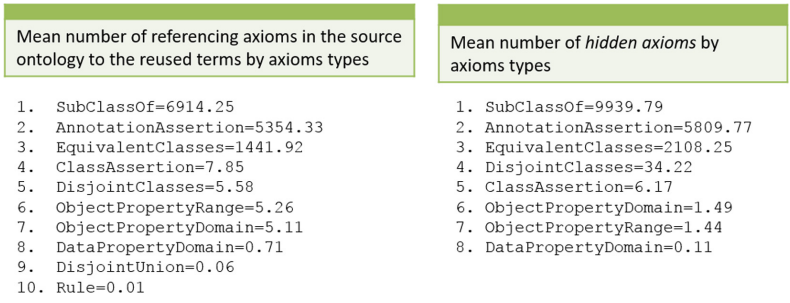

Fig. 6. Ranking of different types of axioms related with the reused terms.

4 Discussion and Conclusions

How much and which content is necessary to reuse is an open discussion in the ontology community. One option is to import (using `owl:import`) the complete *external ontology* when at least one term is reused. This option may require high computational resources when reasoning is required, since even the content that is not reused should be processed by a reasoner. This motivated the development of the MIREOT principle [1] promoting the URIs reuse. MIREOT is likely to be the main reason of the implicit reuse to avoid working with too large ontologies and to not worry about the potential unintended inferences if the complete ontology is imported, what can be criticized from a formal point of view.

The goal of our method is to study the amount of logical knowledge in the *external ontologies* that could be used to axiomatically enrich the *source ontologies*. We have designed a strategy that complements both implicit and explicit reuse. For this reason, we decided to start by analysing the already reused terms. It is worth pointing out that the number of terms shown in Fig. 5 represents less than 2% of all the terms implicitly and explicitly defined in the *external ontologies* (see the graphical representation in our webpage). Increasing the number of terms reused, which could be in line with works such as [7,8], is out of the scope of this work, except for those linked to *hidden axioms*.

Our method requires us to find the ontology to which each reused term belongs. This is currently performed through URI analysis, but this would exclude all the terms that do not follow the URI principle. For example, the URI in the row 3 of Table 1 is quite close to the OBO proposed format; row 4 uses an old reference to the updated term http://purl.obolibrary.org/obo/BFO_0000031. This is a limitation of our current implementation as the method could be improved to use heuristics to overcome such issues or to handle XREF references. Moreover, for all the ontologies associated with terms, the method needs to process their complete implementation. Otherwise, the method could not compute the module or have the information about the potential amount of knowledge that could be reused. Therefore, the results presented here must be contextualised to the set of 123 OBO Foundry ontologies that were successfully processed.

Concerning the impact of our method in current ontologies, the extracted modules could be reused through `owl:import` operations, which would include all the mentioned *hidden axioms*/related terms. This would contribute to the *quality assurance* of *source ontologies* from a logical point of view, and reasoners could use this new content to make inferences. Despite those terms in the modules are selected because they are linked through logical relations with terms already reused in the *source ontology* (used as a *seed signature*), it should be measured if they are conceptually of interest for the *source ontology*. Moreover, once the modules are explicitly imported a reasoner should be used to check the consistency of the enriched ontology. Therefore, our method can be used as a complementary and automatic approach to the application of the MIREOT principle with the Ontofox [12] tool, where ontology developers manually configure what to import using, e.g., a SPARQL-based ontology term retrieval algorithm.

In conclusion, we believe that our method contributes to the *quality assurance* of biomedical ontologies. The paper describes the application of the method to characterise the reuse within the OBO Foundry ontologies. This corpus has been selected because this community builds ontologies by applying a set of shared principles. The findings are that 49 ontologies do not make any type of reuse and 75 do reuse terms. Implicit reuse is the predominant action, that being complemented in a 50% of the cases with explicit reuse. The study of the reused terms has permitted us to visualise the dependencies between ontologies and to cluster them according to the number of ontologies and terms reused. Finally, the exploration of axioms that reference to already reused terms, has revealed that the combination of the content currently being reused with our modular extraction strategy might contribute to increase the axiomatic content of current ontologies, with both new terms and axioms. As future work, we propose the analysis of a larger set of ontologies, improving the mechanism for linking terms with the *source ontology*, and studying the impact of axiomatic richer ontologies in tools that exploit the semantics of biomedical ontologies like [13].

Acknowledgements. This work has been partially funded by to the Spanish Ministry of Economy, Industry and Competitiveness, the FEDER Programme and by the Fundación Séneca through grants TIN2014-53749-C2-2-R and 19371/PI/14.

References

1. Courtot, M., Gibson, F., Lister, A.L., Malone, J., Schober, D., Brinkman, R.R., Ruttenberg, A.: MIREOT: the minimum information to reference an external ontology term. Appl. Ontol. **6**(1), 23–33 (2011)
2. Delbru, R., Tummarello, G., Polleres, A.: Context-dependent OWL reasoning in sindice - experiences and lessons learnt. In: Rudolph, S., Gutierrez, C. (eds.) RR 2011. LNCS, vol. 6902, pp. 46–60. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23580-1_5
3. Ghazvinian, A., Noy, N.F., Musen, M.A.: How orthogonal are the OBO foundry ontologies? J. Biomed. Semant. **2**(2), S2 (2011)
4. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: Modular reuse of ontologies: theory and practice. J. Artif. Int. Res. **31**(1), 273–318 (2008)
5. Horridge, M., Bechhofer, S.: The OWL API: a JAVA API for OWL ontologies. Semant. Web **2**(1), 11–21 (2011)
6. Jiménez-Ruiz, E., Grau, B.C., Sattler, U., Schneider, T., Berlanga, R.: Safe and economic re-use of ontologies: a logic-based methodology and tool support. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 185–199. Springer, Heidelberg (2008). doi:10.1007/978-3-540-68234-9_16
7. Kamdar, M., Tudorache, T., Musen, M.A.: A systematic analysis of term reuse and term overlap across biomedical ontologies. Semant. Web, 1–19 (2016). http://content.iospress.com/articles/semantic-web/sw238
8. Quesada-Martínez, M., Mikroyannidi, E., Fernández-Breis, J.T., Stevens, R.: Approaching the axiomatic enrichment of the gene ontology from a lexical perspective. Artif. Intell. Med. **65**(1), 35–48 (2015)

9. Simperl, E.: Reusing ontologies on the semantic web: a feasibility study. Data Knowl. Eng. **68**(10), 905–925 (2009)
10. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25**(11), 1251–1255 (2007)
11. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. **39**(Suppl 2), W541–W545 (2011)
12. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Res. Notes **3**(1), 175 (2010)
13. Znaidi, E., Tamine, L., Latiri, C.: Answering PICO clinical questions: a semantic graph-based approach. In: Holmes, J.H., Bellazzi, R., Sacchi, L., Peek, N. (eds.) AIME 2015. LNCS, vol. 9105, pp. 232–237. Springer, Cham (2015). doi:10.1007/978-3-319-19551-3_30