HHS Public Access

Author manuscript

Semant Web. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

Semant Web. 2017 ; 8(6): 853–871.

A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies

Maulik R. Kamdar*, **Tania Tudorache**, and **Mark A. Musen**

Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University

Abstract

Reusing ontologies and their terms is a principle and best practice that most ontology development methodologies strongly encourage. Reuse comes with the promise to support the semantic interoperability and to reduce engineering costs. In this paper, we present a descriptive study of the current extent of term reuse and overlap among biomedical ontologies. We use the corpus of biomedical ontologies stored in the BioPortal repository, and analyze different types of reuse and overlap constructs. While we find an approximate term overlap between 25–31%, the term reuse is only <9%, with most ontologies reusing fewer than 5% of their terms from a small set of popular ontologies. Clustering analysis shows that the terms reused by a common set of ontologies have >90% semantic similarity, hinting that ontology developers tend to reuse terms that are sibling or parent–child nodes. We validate this finding by analysing the logs generated from a Protégé plugin that enables developers to reuse terms from BioPortal. We find most reuse constructs were 2-level subtrees on the higher levels of the class hierarchy. We developed a Web application that visualizes reuse dependencies and overlap among ontologies, and that proposes similar terms from BioPortal for a term of interest. We also identified a set of error patterns that indicate that ontology developers did intend to reuse terms from other ontologies, but that they were using different and sometimes incorrect representations. Our results stipulate the need for semi-automated tools that augment term reuse in the ontology engineering process through personalized recommendations.

Keywords

Descriptive Study; Ontologies; Biomedical Domain; Term Reuse; Term Overlap; Composite Mappings; Visualization

1. Reuse in biomedical ontologies

The biomedical research community has been one of the earliest adopters of ontologies to tackle the challenges of efficient knowledge organization, optimized information retrieval and effective annotation of datasets. Researchers have used ontologies for various purposes such as knowledge management, semantic search, data annotation, data integration,

*Corresponding author. maulikrk@stanford.edu.

Editor(s): GQ Zhang, Case Western Reserve University, USA

Solicited review(s): Zhe He, Florida State University, USA

Open review(s): Licong Cui, Case Western Reserve University, USA

exchange, decision support and reasoning [1,2]. For example, *i)* the National Cancer Institute Thesaurus (NCIT) has been used as a reference terminology for cancer data [3], *ii)* the Gene Ontology (GO) has been ubiquitously used for enrichment analysis on gene sets obtained from microarray experiments [4], and *iii)* the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) has been used for the electronic exchange of clinical health information [5].

Over the years, ontology development has become a reuse-centric process [6,7]. All methodologies strongly encourage reuse while building new ontologies, be it at the level of an ontology, or at the level of individual terms [8,9]. In the literature, we may find two areas that benefit from reuse: *i)* ontology engineering, in which experts can reuse already existing ontology structures, and thus reduce the engineering costs; and *ii)* ontology application, in which reuse supports the semantic interoperability among different datasets and applications. For example, the 11th revision of the International Classification of Diseases (ICD-11) reuses terms from SNOMED CT to support its use in electronic health records [10,11]; while federated search engines benefit from reuse by being able to query multiple, heterogeneous knowledge sources without the need for extensive ontology alignment [12].

Several large, collaborative efforts are trying to streamline the development of interoperable, logically well-formed and accurate biomedical ontologies. They deal with ontological term overlap and reuse in different ways. For example, one of the key aims of the Open Biological and Biomedical Ontologies (OBO) Foundry [13] is to create a set of *orthogonal* ontologies by: *i)* defining each term in exactly one ontology, and referring it in other ontologies using its Internationalised Resource Identifier (IRI), or *ii)* using the *xref* mechanism to create references between similar terms in different ontologies [14]. Another prominent example is the Unified Medical Language System–UMLS [15], which uses the notion of a Concept Unique Identifier (CUI) to map terms with similar meaning in different terminologies *a posteriori*. Figure 1 shows examples for the different types of reuse (IRI, CUI and *xref*) employed by various ontology development projects.

For the purpose of this work, we define a **term** to be a class in an ontology. A term usually has a preferred label, other labels, synonyms, and other properties. We define as **term reuse** the situation in which the same term is present in two or more ontologies either by the direct use of the same IRI, or via explicit references (*xref*) and mappings (CUI). We further classify the reuse: (1) *reuse of an ontology*, through the means of the import mechanism available in OWL [16], meaning that the entire source ontology is imported into the target ontology; and (2) *reuse of terms* from one source ontology into another. In many cases, experts reuse not only one term from one ontology, but rather subsets of terms from multiple ontologies (e.g., subtrees). We define as **term overlap** the situation in which two terms are similar, when compared using their labels or synonyms. If we subtract from the set of all overlap terms the reused ones (*term overlap–term reuse*), we will get a set of terms that could have been reused potentially, but have not been in practice. We call this set the **overlap–reuse gap**. Ideally, we should try to minimize this gap.

For this research, we use the entire set of biomedical ontologies stored in BioPortal [17], an open content repository of biomedical ontologies and terminologies. The key contributions of this research can be described as follows:

1. We provide a systematic study of the current state of reuse and overlap across biomedical ontologies.
2. We propose and implement a new approach to determine term overlap across ontologies using composite mappings.
3. We develop a clustering method to help identify patterns of reuse using semantic similarity among ontology terms, and validate the results using the BioPortal Import Plugin logs.
4. We implement a Web application that can search for similar and reused terms in Bioportal ontologies, and that can visualize reuse dependencies and overlap among ontologies.
5. We discuss the state and challenges of reuse in biomedical ontologies.

All results of this paper, as well as all developed visualization tools, are available online at: http://onto-apps.stanford.edu.

The paper is structured as follows: Section 2 describes the related work to this research. Section 3 presents the methods that we used for our descriptive study. Section 4 details the results of applying the research methods, and then we discuss our findings in Section 5.

2. Related Work

2.1. Benefits and challenges of reuse

Ontology reuse is recommended in the methodologies and guidelines outlined by several engineering groups as a means to develop modular, interoperable, accurate and cost-effective ontologies [6,18,19]. Bontas et al. [20] provide several realworld use cases for the benefits of ontology reuse in biomedicine and eRecruitment. By empirically analyzing methodologies, methods and tools currently used, Simperl et al. [7] identify the research and development challenges for ontological knowledge reuse to become a feasible alternative to other ontology-development strategies. In essence, reuse can be increased through the development of pragmatic methods and semi-automated tools that optimally exploit human and computational intelligence for reusing ontologies through a context- and task-sensitive approach [7]. Ontology modularisation techniques (i.e., extracting parts of an ontology using some structural or logical properties) are also an important factor in supporting reuse. Researchers have undertaken comprehensive studies of existing modularization techniques [21,22].

2.2. Tools to support reuse

There are only a few tools that support term reuse in biomedical ontologies. OntoFox [23] is a Web-based application that allows users to retrieve terms, selected properties, and annotations from the source ontologies, using MIREOT principles [24]. The BioPortal Import Plugin [25,26] is an extension of the Protégé ontology editor [27] that allows the

importation of terms, their properties and class subtrees from BioPortal ontologies. The MIREOT Protégé Plugin [28] and DOG4DAG [29] are also Protégé plugins that provide term importations from external ontologies. ProtégéLov [30] allows reuse of terms from the Linked Open Vocabularies repository [31] using `owl:equivalentClass` and `rdf:subClassOf axioms`. All these tools require the users to have prior knowledge of the ontologies where their desired term of interest exists.

2.3. Previous analyses of reuse and overlap

Matentzoglu et al. [32] provide a method to analyze the overlap between automatically-downloaded OWL ontologies from the Web. Ontologies with 90% overlap or containment relations were considered similar. Poveda et al. [33] analyzed the landscape of reuse in the ontologies referenced in Linked Open Data (LOD). The results indicate that over 40% of the terms are reused from other vocabularies, 67% of which are reused by imports, and the rest by referencing the term IRI.

In 2010, a systematic analysis of the member and candidate ontologies in the OBO Foundry indicated that the OBO Foundry had made significant progress over a period of two years towards the goal of orthogonality [34]. However, *term overlap* — percentage of similar terms between the OBO Foundry ontologies, also increased [34].

Five years later, we conducted a study [35] to investigate the level of reuse across all the biomedical ontologies stored in BioPortal [17]. Both these studies carried out simple lexical comparisons of the term labels to determine term overlap. Even though effective, this naive method tends to leave out terms that represent the same concept but have lexically–different term labels (e.g., `“Cardiac Muscle”` and `“Myocardium”`). For the three types of reuse observed in the biomedical domain (IRI, *xref* and CUI, Figure 1), we estimated term reuse using a simple metric (Equation 1).

$$Reuse = \frac{\text{unique reused terms}}{\text{total terms}} \quad (1)$$

We found term reuse to be 3.1%, 3.9% and 4.1% for the three reuse types respectively, whereas, we found a term overlap of 14.4%. We also found that most ontologies reuse less than 5% of their terms. These terms are reused from a small set of popular ontologies only. We presented some use cases, in which the developers reused terms with different, and often, incorrect representations.

In this paper, we will extend this research by providing a new approach to determine term overlap, a better metric to estimate term overlap and reuse, and a deeper understanding of how ontology developers reuse terms.

3. Methods

For our descriptive study, we employed several methods that aim to: (i) estimate the level of term reuse and term overlap across biomedical ontologies, (ii) extract reuse patterns from

BioPortal ontologies, and (iii) extract reuse patterns from time-stamped BioPortal Import Plugin logs. These methods are inspired from text mining, graph theory and unsupervised learning. We make the results available through interactive visualizations and a search application (http://onto-apps.stanford.edu). Figure 2 describes the workflow of our methodology and the methods used stepwise. The structure of this section follows the numbered steps of the workflow.

3.1. Datasets

We used two datasets for our study: (i) a dump of BioPortal ontologies to analyse term reuse (Step 2) and overlap (Step 3), as well as to perform the clustering (Step 4); and (ii) the logs of the BioPortal Import Plugin to analyze the patterns of reuse in user ontologies (Step 5).

3.1.1. BioPortal ontologies—We obtained a triplestore dump of the BioPortal ontologies in N-triples format that contained 509 distinct ontologies as of January 1, 2015. This dump did not contain some ontologies that were deprecated or merged with existing ontologies, or added to BioPortal after January 1, 2015. After removing ontological views (i.e. $\mathcal{O}_1 \subseteq \mathcal{O}_2$), we were left with 377 distinct biomedical ontologies (Figure 2, Step 1). These ontologies include 8 OBO Foundry member ontologies (GO, CHEBI, PATO, OBI, ZFA, XAO, PR and PO), 105 OBO Foundry candidate ontologies (e.g., OGMS, HP) and 31 UMLS Terminologies (e.g., SNOMED CT, ICD-9).

3.1.2. BioPortal Import Plugin logs—The BioPortal Import Plugin, an extension to the Protege ontology editor, allows users to import terms and sub-trees from BioPortal ontologies into their own ontology [25,26]. The plugin invokes the BioPortal REST API to search the BioPortal ontologies, and also to import terms.

We obtained the logs of REST calls that the plugin made to BioPortal. The logs are time and IP-stamped, and span the period from 26th September, 2011 – 14th May, 2013 (~20 months). Listing 1 shows an excerpt of these logs.

Even though we did not have access to the user ontologies into which these imports were performed, these logs were an important source of information of terms that were reused together in user ontologies. We used these logs to identify patterns of reuse (Figure 2, Step 5).

3.2. Identifying Term Reuse

For the purpose of this work, we define as **term reuse** the situation in which the same term is present in two or more ontologies, either by the direct use of the same IRI, via explicit *xref* references, or via CUI mappings.

To identify term reuse (Figure 2, Step 2), we used the BioPortal corpus (Section 3.1.1), and defined three reuse constructs:

1. *IRI* – two terms share the same IRI,
2. *xref* – two terms are linked through the *xref* annotation [14], and

3. *CUI* – two terms are mapped to the same UMLS CUI.

We iterated over all the axioms in each of the 377 BioPortal ontologies to extract class term IRIs, their labels, synonyms, *xref* links and UMLS CUI mappings, when available. From the 5,718,275 class terms, we used the three constructs (same IRI, *xref* annotation, and CUI mapping) to extract the set of terms that satisfy any of the three reuse criteria (Figure 1). For the first two reuse types (*IRI* and *xref*),[1] we identified the source ontology for each term using a heuristic approach described previously [35]. For each ontology, we calculated:

1. The percentage of terms reused using the first two constructs from other ontologies (IRI and *xref*),
2. The total number of ontologies reused from,
3. The percentage of terms reused by other ontologies,
4. The total number of other ontologies reusing terms,
5. CUI-mapped terms among other ontologies,
6. Reuse among all distinct pairs of ontologies.

Using these metrics, we determined those ontologies that reused the maximum number terms from other ontologies, and also those ontologies whose terms were reused the most.

We generated a graph $\mathscr{G}$, where the terms identified through IRIs represent nodes (Figure 3a). The number of ontologies in which the term is reused is represented as an attribute of each node. An *xref* annotation is shown as a unidirectional arrow, whereas all terms mapped to the same CUI are interlinked with each other using bidirectional arrows. A *component* of a graph is a subgraph in which any two nodes (terms) are connected to each other by paths, and the subgraph is connected to no additional node in the main graph. Due to the nature of these ontological terms (generally distinct for a given ontology), we produced a graph composed of different, disjoint components (e.g., $\mathscr{T}_1$, $\mathscr{T}_2$ and $\mathscr{T}_3$ are different components in Figure 3a). This graph can be divided based on the type of the edges, and thus yields three modules corresponding to our three reuse constructs:

IRI reuse module – the graph module containing only IRI edges (an undirected edge links two terms with same IRI),

***xref* reuse module** – the graph module containing only *xref* edges (a directed edge links the source term and the referenced term via *xref*), and

CUI reuse module – the graph module containing only CUI edges (an edge links two terms that are mapped to the same CUI).

For each reuse module, we calculated the term reuse across all biomedical ontologies using the equation given below, where N represents the total number of terms extracted (5, 718, 275), $\mathscr{M}_r$ is a reuse module, composed of k components $\{\mathscr{T}_0, \mathscr{T}_1, \ldots, \mathscr{T}_k\}$. Each component $\mathscr{T}_j$ is formed from n_j terms, i.e. $\{t_{0j}, t_{1j}, \ldots, t_{nj}\} \in \mathscr{T}_j$. The number of terms in a component T_j must follow $1 < n_j < N$ (i.e., components with a single term are not allowed). All terms in

[1] UMLS CUI reuse was excluded, as we could not identify the source ontology for a CUI.

one component are reused forms for the same term. We calculate term reuse for each of the three different reuse modules:

$$Reuse = \frac{\sum_{j|T_j \in \mathscr{M}_r} n_j - k}{N} \quad (2)$$

The above equation serves as a better metric to estimate term reuse as compared to the previous metric (Equation 1). The equation calculates the percentage of terms in BioPortal that are not unique, but are reused, unlike the previous metric, which did not include the count of reused versions of a term.

3.3. Detecting Term Overlap through composite mappings

For the purpose of this work, we define **term overlap** as the situation in which two terms are similar, when compared using their labels or synonyms. To detect term overlap (Figure 2, Step 3), we use the BioPortal corpus (described in Section 3.1.1).

In our initial approach [35], we normalized the term labels by converting them to lowercase and then removing all non-alphanumeric characters. We performed naïve string matching to determine the potential *term overlap*. However, we realized that the terms with labels such as `"Cardiac Muscle"`, `"Heart Muscle"`, `"Muscle of Heart"` and `"Myocardium"` would be treated as separate terms in this approach, when these terms are the same and should be treated as term overlap.

To overcome this limitation, we considered using composite mappings in the current approach. Given a mapping from $A \rightarrow B$ and from $B \rightarrow C$, where terms $A \in \mathscr{O}_1, B \in \mathscr{O}_2$ and $C \in \mathscr{O}_3$ and $\mathscr{O}_1, \mathscr{O}_2, \mathscr{O}_3$ are different ontologies, a mapping from $A \rightarrow C$ is called a *composite mapping* [36]. This approach, which leverages transitivity of terms, has been used in the past to match unstructured vocabularies using a background ontology, where $\mathscr{O}_2$ is a background ontology [37]. An example of such a composite mapping is shown in the table below.

We extended this notion to generate graphs of such composite mappings ($\mathscr{M}$) between different terms across all BioPortal ontologies, without predefining any particular ontology as a background ontology. We extracted preferred labels ($\mathscr{L}$), exact synonyms ($\mathscr{S}_{\mathscr{E}}$), related synonyms ($\mathscr{S}_{\mathscr{R}}$), and other synonyms ($\mathscr{S}_{\mathscr{O}}$) from the sources listed in Table 2.

We normalized the labels and synonyms, by first removing a set of 126 common English stop words (e.g. *"of"*), and then converting them to count vectors. We calculated cosine similarities between each pair of these string vectors and established a mapping, if the similarity was $>95\%$. Due to the size and relative reduced importance of $\mathscr{S}_{\mathscr{O}}$, we also considered bi-gram phrases of words in similarity calculations.

We generated 5 different overlap modules from different combinations of composite mappings:

1. $\mathscr{LG}: \{\forall m \in \mathscr{LL}\}$
2. $\mathscr{LEG}: \{\forall m \in \mathscr{LL} \cup \mathscr{LS}_{\mathscr{E}} \cup \mathscr{S}_{\mathscr{E}}\mathscr{S}_{\mathscr{E}}\}$
3. $\mathscr{LERG}: \{\forall m \in \mathscr{LL} \cup \mathscr{LS}_{\mathscr{E}} \cup \mathscr{LS}_{\mathscr{R}} \cup \mathscr{S}_{\mathscr{E}}\mathscr{S}_{\mathscr{E}} \cup \mathscr{S}_{\mathscr{E}}\mathscr{S}_{\mathscr{R}} \cup \mathscr{S}_{\mathscr{R}}\mathscr{S}_{\mathscr{R}}\}$
4. $\mathscr{LERCG}: \{\forall m \in \mathscr{LL} \cup \mathscr{LS}_{\mathscr{E}} \cup \mathscr{LS}_{\mathscr{R}} \cup \mathscr{LS}_{\mathscr{C}}\}$
5. $\mathscr{XG}: \{\forall m \in \mathscr{M}\}$

The $\mathscr{LG}$ overlap module contains only the mappings performed using the properties from the $\mathscr{L}$ set defined in Table 2 (that is, `skos:prefLabel`, `rdfs:label`, `dc:title`). The $\mathscr{LEG}$ overlap module includes besides the *label–label* mappings, also the *label–exact synonym* and *exact synonym–exact synonym* mappings.

The final $\mathscr{XG}$ overlap module contains all the composite mappings in $\mathscr{M}$. We removed the edges that were present in the three reuse modules from $\mathscr{LEG}$ (i.e., overlapping terms that were already reused), to find the *overlap–reuse gap*. This new module is called $\mathscr{LEG} - \{Reuse\}$, where {*Reuse*} = {*IRI*} ∪ {*xref*} ∪ {*CUI*}. The $\mathscr{LERCG}$ overlap module and $\mathscr{LEG} - \{Reuse\}$ module are shown in Figure 3b and c.

In the next step, we identified those terms that had the same source ontology and identifier, but a different IRI representation, and no explicit mappings (e.g., `OBO:owlapi/fma#FMA_31396` was used instead of `OBO:FMA_31396`). Such situations show that ontology developers intended to reuse a term, but they used different, and sometimes incorrect term representations. These situations do not represent actual reuse, and we marked such cases as *intent for reuse* (Section 5). We removed any interconnecting edges between terms that show an *intent for reuse* in $\mathscr{LEG} - \{Reuse\}$ to generate the final module $\mathscr{LEG} - \{Reuse, Intent\}$.

We calculated term overlap for each overlap module using the metric described in Equation 2, where all nodes (terms t_{ij}) in $\mathscr{T}_j$ (connected component of composite mappings) of the overlap module $\mathscr{M}_o$ can be considered singular (Figure 3).

For each the five overlap modules, we conducted an empirical analysis on the composition of the term labels of 100 randomly selected components to determine the threshold of the maximum distance (mapping hops) between two leaf nodes, for which any component $\mathscr{T}_j$ can be considered to be *'pure'* (i.e., contains terms that can still be considered similar). We identified the maximum distance (i.e., mapping hops) for which the components are still 'pure' to lie between [8,10], depending on the overlap module.

We called the components that have mappings exceeding the maximum distance *Hybrid Components*. These components are "hybrid" because they contain terms that are likely not similar to each other, usually because of a faulty mapping. In essence, the hybrid components can also be broken down into smaller components that are joined by one incorrect edge caused by a faulty mapping. Term nodes in these smaller components may be similar to each other. In the example from Table 3, term t_3 has a faulty synonym

`Intercalated disk` that links two smaller, relevant components $\mathscr{T}_{1a}$ and $\mathscr{T}_{1b}$ creating a hybrid component $\mathscr{T}_1$.

We calculated another term overlap estimate, which we called *Non-hybrid Term Overlap*, by excluding hybrid components from consideration in our metric. By excluding hybrid components altogether from this estimate, we set a lower bound on our estimated term overlap.

3.4. Clustering to detect patterns of reuse

One goal of this work is to investigate whether the reuse within biomedical ontologies occur in certain patterns that can be identified algorithmically. To this end, in Step 4 of our workflow (Figure 2), we used a two-phase clustering approach on the IRI module that we defined in Section 3.1.1. As a reminder, the IRI reuse module contains only IRI edges that link terms that share the same IRI.

We excluded the *CUI* and *xref* reuse modules from this analysis, as *CUI* mappings and *xref* annotations are generally established *a posteriori* in the engineering process.

Using the terms in the *IRI* reuse module, we generated a term–ontology matrix. The rows contain the terms that have been reused at least once (i.e., the term appears in at least 2 ontologies with the same IRI), and the columns contain the ontology in which the term appears. Whether a term exists in an ontology or not was indicated as 1 or 0 respectively, resulting in a very large, sparse, binary matrix.

As our term–ontology matrix X is categorical (n terms, m ontologies), we used a *k–modes* algorithm [38] over 100 simulations with different k to partition the terms into large, disjoint clusters (k). The initial step is similar to the the *k–means* algorithm, where k unique terms are selected as cluster centroids $Z = \{Z_1, Z_2, \ldots, Z_k\}$. *k–modes* algorithm assigns a term X_i to a cluster whose centroid Z_l has the minimum distance $d(X_i, Z_l)$ to it. $\delta(x_j, z_j)$ checks if the term and the cluster centroid are present/absent together for one ontology $\mathscr{O}_j (\therefore \delta(x_j, z_j)=0)$. After each term is assigned to a cluster, new centroids are generated for each cluster based on the modes of values for each ontology $\mathscr{O}_j$ (i.e. if more terms in a cluster Z_l are present in $\mathscr{O}_j$ then $z_{l,j} = 1$). Until the cluster centroids are stable, we iterated over these steps. Over 100 simulations, the value of k is chosen with a desirable measure of cluster compactness (minimum spread of each cluster) and separation (maximum distance between cluster centroids).

$$\delta(x_j, z_j) = \begin{cases} 0 \text{ if } (x_j = z_j) \\ 1 \text{ if } (x_j \neq z_j) \end{cases} \quad (3)$$

$$d(X_i, Z_l) = \sum_{j=1}^{m} \delta(x_{i,j}, z_{l,j}) \quad (4)$$

For each pair of terms in each cluster, we computed a similarity score as follows:

$$Sim(A,B)=\omega_1\left(\frac{|\mathscr{O}_A\cap\mathscr{O}_B|^2}{|\mathscr{O}_A\cup\mathscr{O}_B|}\right)+\omega_2\left(\frac{|\mathscr{S}\mathscr{P}_A\cap\mathscr{S}\mathscr{P}_B|}{|\mathscr{S}\mathscr{P}_A\cup\mathscr{S}\mathscr{P}_B|}\right) \quad (5)$$

In the equation above, $\mathscr{O}_A\cap\mathscr{O}_B$ indicates the set of common ontologies between terms A and B. $\mathscr{S}\mathscr{P}_A=\{x|x\supseteq A\}$, and $\mathscr{S}\mathscr{P}_A\cap\mathscr{S}\mathscr{P}_B$ indicates the set of common super terms of A and B.

As can be seen, the similarity measure is a weighted distribution of common ontologies and Jaccard semantic similarity. $\omega_1 > \omega_2$, as we want to discern how ontology developers reused terms based on the set of ontologies in which these terms co-occur. We consider the proportion of shared terms, to reduce the impact of `owl:Thing` and other upper-level ontology terms which would be reused in many ontologies.

We generated a term–term affinity matrix A, where $A_{ij} \geq 0$ represents the similarity between the terms i and j. We used Spectral Clustering [39] over this matrix to further partition each large cluster. This method uses the largest eigenvectors of the similarity matrix to perform dimensionality reduction before using *k–means* clustering in the fewer dimensions. We performed 100 simulations with different values of ω_1 and ω_2 to isolate sub-clusters that are composed of terms from one source ontology only. Based on the current state of tools that support reuse, as well as the mental processing of the ontology developers, terms or groups of terms reused together in one session originate from the same source ontology.

3.5. Analyzing BioPortal Import Plugin Logs

In step 5 of our workflow (Figure 2), we analyzed the logs generated by the BioPortal Import plugin (see Section 3.1.2). We used this analysis for two purposes: (1) to gain knowledge on other reuse patterns that occur in user ontologies, and (2) to validate whether the insights generated from our clustering analysis are accurate.

The entries in the BioPortal logs are generated as the user does certain operations in the user interface of the plugin. For example, if the user searches for a term in a BioPortal ontology using the plugin, the log will record a line corresponding to the *search* REST call made to BioPortal (see Listing 1). An import operation in the plugin would trigger other REST calls.

As we do not have access to the user ontologies into which the BioPortal terms have been imported, the only sources we have are the time- and IP-stamped BioPortal call logs. Therefore, we had to reverse-engineer these logs to find out the actions that the users have taken in the user interface, and to identify which BioPortal terms are being reused (i.e., imported) together.

We documented the algorithm we used to reverse-engineer the logs in the additional online materials (http://onto-apps.stanford.edu).

As a result of running the reverse-engineering algorithm on BioPortal logs, we obtained term sets that have been reused (i.e., imported) together in user ontologies. Then, we mapped the extracted terms to existing terms in the current version of the source BioPortal ontology to find the overall depth of tree imports and the location of these terms and subtrees. We used this information as an additional source of reuse patterns, and also to validate the hypotheses made from clustering analysis (Section 3.4).

4. Results

We now present the results of each of the methods that compose our workflow (Figure 2), described previously in Section 3.

4.1. Reuse

Previously, we found that most ontologies reuse less than 5% of the total terms in their current versions, using either the same IRI or through *xref* annotations [35]. Out of 377 BioPortal ontologies, 156 did not reuse any term using the IRI construct, and 315 did not reuse through *xref*. Moreover, ontologies reused terms from a small set of popular ontologies only. More than 250 ontologies have no terms reused. Figure 4 shows histograms of the percentage of terms that are reused by other ontologies. We also observed that there are 20 ontologies that exhibit reuse between 95% to 100% of their total terms. These ontologies are developed by reusing combinations of multiple ontologies (e.g., CCONT reuses terms from EFO, NCBITAXON, ORDO, and 19 other ontologies).

Using our *CUI* construct, we found: *i)* popular UMLS terminologies such as ICD10CM (ICD10 -Clinical Modification), LOINC (Logical Observation Identifiers Names and Codes), HL7 (Health Level Seven Reference Implementation Model, Version 3) and MESH (Medical Subject Headings) to be composed primarily of unshared, unique terms, *ii)* procedural terminologies such as HCPCS (Healthcare Common Procedure Coding System), CPT (Current Procedural Terminology) and ICD10PCS (ICD10 – Procedure Coding System) have very few terms mapped to the same CUI, and *iii)* Several new terms were introduced in ICD10CM during the migration from ICD9CM, potentially impacting reuse [35].

The 16 ontologies whose terms are reused the most from the first 2 constructs (IRI and *xref*) are shown in Figure 5. The plot indicates the number of ontologies (#) that reuse terms from a given ontology as dots, and the percentage of terms (%) that are reused with respect to the number of terms in their current version as bars. For example, 95.2% of the total terms in the current version of GO are reused using the same IRI by 74 ontologies. Also, 3.7% of the total GO terms are *xref*-linked in 37 ontologies.

It is easily noticeable that most of these are popular or upper-level ontologies, some of which have more than 100% of their terms reused (e.g., we found 101 different versions of Basic Formal Ontology – BFO IRIs, whereas the current version only has 39 terms). As we have discussed [35], this anomaly is due to the fact that ontology developers tend to reuse terms with different versions, notations, or namespaces, that are sometimes incorrect and

have no explicit mappings to the original term. We do not consider this case as reuse, but rather an **intent for reuse**, and we discuss it in Section 5.

Using the updated metric described in Section 3.2, we found term reuse to be 6.63% for the *IRI* reuse module, 5.98% for the *xref* reuse module, and 8.39% for the *CUI* reuse module.

4.2. Overlap

4.2.1. Term Overlap—In our previous work [35], we determined term overlap using a naive approach. We found a total of 2, 023, 854 terms sharing 752, 177 unique labels across the BioPortal ontologies. Using the new metrics described in Section 3.3, we can calculate this naive term overlap to be 22.23%. In addition, the new metrics allowed us to compute more precise overlap statistics that we show in Table 4.

The $\mathscr{LG}$ module is the most similar to our previous naive term overlap method, as this module contains only mappings $\forall m \in \mathfrak{LL}$ (label–label mappings). However, there is a substantial increase in the level of the term overlap from 22.23% to 25.37% (non-hybrid term overlap).

Once we include also the other types of mappings using synonyms (rows 2–6 in Table 4), the term overlap gradually increases all the way up to 32.75%, although the number of hybrid components also increases. It is noteworthy to see that the non-hybrid term overlap is almost similar to the term overlap of $\mathscr{LG}$ module (≈25%).

Rows 6 and 7 in Table 4 show that after removing all the three reuse modules (cf. Section 3.3), the term overlap decreases—the range is (18.21%, 21.62%). On evaluating the $\mathscr{LEG} - \{Reuse, Intent\}$, we find that the term overlap drops down to (13.21%, 16.57%). Obviously, this term overlap statistic captures only the intent for reuse rather than actual reuse.

4.2.2. Ontology Overlap—As a next step, we investigate how the term overlap reflects on ontology overlap. Therefore, we mapped the nodes in the $\mathscr{LEG} - \{Reuse\}$ module to their respective ontologies, and created an edge between all the pairs of ontologies, if there existed an edge between the nodes (i.e., $\forall e = (n_1, n_2)$, *s.t.* $e \in \mathscr{LEG} - \{Reuse\}$, $n_1 \in \{\mathscr{O}_1, \mathscr{O}_2\}, n_2 \in \{\mathscr{O}_3\} \Rightarrow \{e(\mathscr{O}_1, \mathscr{O}_3), e(\mathscr{O}_2, \mathscr{O}_3)\}$). After removing all the terms and aggregating all edges between two ontology nodes to a single edge with a weight $w = \Sigma e$, we have an undirected ontological overlap graph with edges depicting the term overlap between two ontologies.

We generated a directed sub-graph (Figure 6) between those ontologies that have more than 30% term overlap with respect to any one of the connected ontologies. Note that, for simplicity, Figure 6 only includes the OBO Foundry member and candidate ontologies (blue squares), UMLS terminologies (red circles), and a few popular ontologies in BioPortal (green octagons). If we were to include all the ontologies in this graph, it would have created an indecipherable visualization. The interactive visualization is available in the online materials (http://onto-apps.stanford.edu).

Figure 6 shows that there is substantial overlap among ontologies generated independently through the OBO Foundry and UMLS methodologies. The overlap between BFO and the OBO Foundry candidate ontologies is caused by the fact that the candidate ontologies import BFO as their upper-level ontology, but they use different (incorrect) IRI representations. It is also noteworthy to see that the UMLS terminologies for adverse events, namely World Health Organization Adverse Reaction Terminology (WHO-ART), Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), and the Medical Dictionary for Regulatory Activities (MEDDRA), have substantial term overlap. The lower region of the graph shows several anatomical ontologies (CARO, UBERON, XAO, TAO, FMA, MA, TGMA, etc.), in which term overlap is obvious (similar anatomical features), but is debatable—most terms represent anatomical parts that may not be necessarily equivalent, as they belong in different organisms. Finally, the top-right corner shows the overlap between the RxNorm Vocabulary and the Drug Ontology (DRON). These results and the **intent for reuse** are described in detail in Section 5.

4.3. Clustering

The first step of our two-phase clustering approach was to use a *k-modes* algorithm over simulations for $k = 2 \rightarrow 100$. We computed cluster compactness and separation by computing the cosine distance between the set of ontologies in one cluster against another. The desired cluster compactness and separation value was found to be at $k = 6$, after which we would have overlapping clusters, or clusters with single terms.

The primary ontological composition of the clusters was determined from the ontologies common among terms in a cluster, and is shown in Table 5. It should be noted that *IRI* reuse was rarely found in UMLS terminologies with the exception of NCBITAXON, NCIT, and SNOMED CT. The primary ontological composition of the terms in the large clusters either consists of: *i)* ontologies that frequently reuse terms from one major source ontology (e.g. CHEBI, GO, NCIT, DOID) in that cluster, or *ii)* one main ontology that reuses terms from multiple other ontologies and exhibits >90% reuse, e.g. CCONT.

We computed an affinity matrix among all pairs of terms in a given cluster using weights $\omega_1 = 0.85$, $\omega_2 = 0.15$. These values were again generated after a set of 100 simulations, so that most of these sub-clusters are generally composed of individual source ontologies.

After executing spectral clustering using the affinity matrix, we divided all the term pairs in each sub-cluster in 2 bins, based on their Jaccard semantic similarity measure (<0.9 in Bin 1, and >0.9 in Bin 2). We plotted the proportion of term pairs in each bin for each cluster. Cluster 4 is shown in Figure 7. In Cluster 4, a larger proportion of term pairs in any given sub-cluster have a semantic similarity in the range of (0.9–1.0) (> 70%), indicating that these are either sibling terms or one term is the direct superclass of another. Generally, we found this to be the case for all the large clusters of the first kind. This finding likely indicates that ontology developers reusing terms from one main source ontology tend to reuse hierarchical subtrees mainly composed of terms with parent–child or sibling relations. This was less evident in the second kind of the large clusters where the proportion ranged between 30–60% of term pairs.

We mapped these sub-clusters to their location in the source ontology. We found that most of these 2-level substructures are located in the higher or upper-middle levels of the ontology. Hence, developers reuse terms from the higher levels in the ontological hierarchy of a small set of popular ontologies, and seldom reuse leaf nodes.

4.4. BioPortal Import Plugin Log Analysis

We found a total of 3,538 distinct IP addresses originating from 90 different countries, from which ontology developers used the BioPortal Import Plugin to search and reuse terms from BioPortal ontologies. We were able to isolate 5,755 individual terms and 2,139 ontological subtrees imported from 40 different ontologies in 516 distinct sessions. For an IP address, a **session** indicates the time period that has no intermittent breaks of >1 hour between two REST API calls. We found a total of 195,894 terms that users imported using the plugin. Out of these, we were able to map 193,601 terms to terms in the current versions of the Bio-Portal ontologies. The remaining terms were either deprecated, or terms such as, `owl:Thing` and `time#datetimedescription` that do not have a designated source ontology.

The top 10 ontologies with the maximum number of sessions were SNOMEDCT, NCIT, BFO, ABA-AMB, FMA, GO, RCD, AMINO-ACID, HP and IAO, whereas with the maximum number of terms were in ICD10PCS, SNOMEDCT, NCIT, ICD9CM, LOINC, BIRNLEX, ABA-AMB, FMA, RCD and SHR.

The ontologies that were reused the most through the plugin, both by the maximum number of sessions or by the maximum number of terms, are shown in Figure 8. The total number of sessions observed, total number of single term imports, total number of structures imported, and total number of terms imported are shown as a bar plot. The structure of the content imported from each source ontology is shown across the depth of an ontology — the imported structures are shown as translucent blue polygon and the terms imported (either single or as a group) are shown as circular constructs, grouped according to the level. The depth of the ontology was retrieved from BioPortal repository. The width of the structure on each level is indicative of the number of terms imported on that level in log scale. The radius of the circular construct represents the total number of terms on that level. For clarity purposes, we have only shown 4 ontologies — FMA, ICD10PCS, NCIT and SNOMEDCT. The website (http://onto-apps.stanford.edu) contains interactive versions of these plots with 16 different ontologies.

In general, we found that, on an average more people tend to reuse terms from OBO Foundry ontologies (higher number of sessions detected) than UMLS terminologies using the Bioportal Import Plugin, with the exception of NCIT and SNOMED CT. However, the users, who import UMLS terminologies, tend to reuse more number of terms, in the form of complete hierarchical structures, during a single import session.

In the cases of ICD10PCS and ICD9CM, we found that the users reuse the entire hierarchy of these ontologies starting from the root node, into their target ontology. We observed the same pattern also in the case of the BFO, but it is expected as it is an upper level ontology. In almost all the other cases, we found that the ontology developers simply reuse terms from

the higher or upper–middle levels in an ontological hierarchy, and the lower leaf nodes and structures are seldom reused. This reuse pattern can be seen in the FMA ontology in Figure 8. We found the same reuse pattern in GO, CHEBI, NCBITAXON and LOINC (http://onto-apps.stanford.edu). As is clearly evident from the SNOMED CT and NCIT, most ontology developers generally import 2–level sub-trees composed of parent–child and sibling terms. These structures are represented as triangular polygons of similar dimensions along the midline of the respective visualizations in Figure 8 with a higher opacity than other structures.

4.5. Reuse and Overlap Visualization on the Web

One of the contributions of our work is a general-purpose visualization of reuse and overlap among biomedical ontologies that employs the reuse and overlap modules, which we generated as part of this work. The Web application also allows users to search for similar terms by providing any string or an IRI as an input. In case of a string, the application matches the name to the set of the most similar terms that have it as a label or a synonym. We believe such an application is of general interest, and we make it available to the community through our website (http://onto-apps.stanford.edu/).

The application does a depth-first search against the $\mathscr{X}\mathscr{G}$ module, and returns all composite mappings, in which each term is a node of. The results are displayed in a tabular, or a force-directed network layout. The interactive force-directed network visualization allows users to explore reuse dependencies and overlap among BioPortal ontologies. Our website also provides access to the module graphs, and the analysis results of the BioPortal Import Plugin logs.

5. Discussion

5.1. Term Reuse

As seen in Figure 4, we are seeing the full spectrum of reuse from 0 – 100%, but in general, reuse is fairly low. Not only do most ontologies in BioPortal never reuse terms, their terms are also never reused by other ontologies, which is contrary to the reference-application paradigm considered in the ontology engineering process. However, we did find some ontologies that are approaching complete reuse. For example, the Mental Functioning Ontology (MF) [40], reuses 91.33% of its terms from 6 different ontologies. Our clustering analysis shows that not only single terms are reused, but also entire hierarchical structures of the source ontologies are reused. Ontology engineers need semi-automated tools to support both cases.

Generally, well-established ontologies and controlled terminologies do not reuse terms from other ontologies. Usually, these ontology are built by large organizations (e.g., NCI, WHO, IHTSDO). Some of these organizations are making concerted efforts to take advantage of reuse. For example, ICD-11 and SNOMED CT are trying to define a common core ontology to be reused by both [11]. Such collaborations may generate a set of best practices for ontology reuse in the future.

Through the empirical analysis of the BioPortal Import Plugin logs, as well as, the generated clusters and overlap modules, we found some reuse patterns that show that ontology developers have the intention to reuse terms. Essentially, these are IRI patterns that generally have the same identifier and source ontology, but that are reused from different versions of the source ontology, or represented using different notations or namespaces. These patterns cannot be considered as term reuse, as the IRIs use different, and often incorrect, representations for the same terms, and no explicit CUI or *xref* mappings were found. Hence, the advantages of term reuse can not be experienced. By using the correct IRI representation, the term overlap could be reduced substantially. We summarize these IRI patterns in Table 6, and provide a few examples for each. We also indicate the recommended representation, where possible.

We found several cases, in which an ontology reuses the same terms from different ontologies, and these terms are not linked by a reuse construct. For example, the BioModels Ontology (BIOMODELS) reuses the same terms from two different ontologies: *i)* `Hepatic Oval Stem Cell` from Cell Ontology (CL) and Foundational Model of Anatomy (FMA), and *ii)* `Xanthopore` from CL and Gene Ontology (GO). Even if these terms are likely equivalent, there is no reuse construct that links them.

Based on the observations from this study that show only modest reuse among biomedical ontologies, we believe that ontology engineers would benefit from better guidelines, along with improved tools, to increase term reuse.

5.2. Term Overlap

In 2010, a systematic analysis of all the OBO Foundry ontologies outlined consistent term overlap, yet minimum term reuse, and commented on the limitations and challenges to achieve *orthogonality* [34]. Five years later, we extended this analysis and estimated term reuse and overlap over the entire continuum of biomedical ontologies (including UMLS terminologies) in the BioPortal repository. We found that we are still very far from achieving desirable term reuse [35]. Most ontologies exhibit considerably less than 5% reuse or no reuse through any constructs, and generally reuse terms from only a small set of ontologies.

The OBO Foundry mandates reuse by candidate ontologies from the member ontologies under its orthogonality aim. However, there is still substantial *term overlap* present among biomedical ontologies, including OBO Foundry ontologies.

In our previous analysis, we used a conservative approach to determine term overlap. As a result, lexically–different terms that may be similar, and can be categorized under term overlap, were considered different. Using our approach of tokenization and removal stop words, we were able to map terms with labels such as `“Muscle of Heart”` and `“Heart Muscle”`, whereas, through different overlap modules of composite mappings from preferred labels and synonyms, we were able to link `“Heart Muscle”`, `“Cardiac Muscle”`, `“Myocardium”`, and also terms in other languages such as `“Myocarde”@FR` and `“Herzmuskel”@DE`. The estimated *term overlap* through these overlap modules ranges from 25%–31.5%.

Our approach for detecting overlap has certain limitations.

1. Terms with labels such as `"Second phalange of the third finger"` and `"Third phalange of the second finger"`, and also `"WAS Gene"` (Wiskott-Aldrich syndrome) and "Gene" will be grouped together — due to count vectors and the exclusion of the stop word *"was"* respectively.
2. Lexically-similar terms in different ontologies may represent different concepts (e.g., anatomical concepts like `spleen` between Zebrafish Anatomy (ZFA) and Xenopus Anatomy (XAO)).
3. Some biomedical ontologies use different classes for the same concept to show evolutionary or developmental stages (e.g. `Myocardium` in Human Development Anatomy, Timed (EHDA) and Abstract (EHDAA) ontologies). We group these classes under term overlap, but they may be different.
4. Some ontologies may instantiate a synonym relation between terms that can actually have an *"is part of"* relation. This choice can lead to false composite mappings (e.g. `Cranium` has the synonyms Skull in the Teleost Anatomy Ontology (TAO)).
5. Some ontologies use chemical formulas as synonyms. Terms with the same chemical formula may be stereoisomeric molecules or completely different compounds (e.g., `(+)-Menthofuran` and `Safranal` ($C_{10}H_{14}O$)). This challenge has also been seen during alignment of different biomedical vocabularies for federated search, where `Aspirin` and `Acetylsalicylic acid` are the same but `L-Glucose` and `D-Glucose` are not the same [41].

Hence, the term overlap estimates should be seen cautiously, and can serve as an upper bound to the actual *term overlap*. Overlapping nodes that are at a path distance of more than 2 edges are generally different, especially if the edges $e \notin \{\mathscr{LL}, \mathscr{LS}_{\mathscr{E}}\}$. To bring these estimates closer to actual overlap, we introduced the concept of bigram similarity for $e \notin \mathscr{S}_{\mathscr{C}}\mathscr{S}_{\mathscr{C}}$ and hybrid components, and the resultant *term overlap* is closer to the one derived from the $\mathscr{LG}$ module.

5.3. Clustering

One of the key challenges that we encountered while clustering was the fact that we were dealing with a large number of terms (compared to the features), resulting in a large $n \times m$ matrix where $n \ggg m$. Also, as the initial matrix consisted only of the IRI-reused term–ontology pairs that are reused on an average between 2–3 ontologies, we had a very sparse binary matrix. There are various methods to deal with this such large, multi-dimensional matrices, ranging from MapReduce [42] to simple candidate generation [43]. Our two-phase approach allowed us to divide the term–ontology pairs into large distinct clusters of terms shared between some common group of ontologies. We could then also include the semantic hierarchy of these terms in the different shared ontologies for a subsequent spectral clustering. We believe that our similarity equation can be extended to incorporate other features such as co-occurrence of these terms in PubMed annotations, and our generated

term–term affinity matrix can be used in a item-based collaborative filtering method to generate recommendations for reuse.

From clustering, we claim the following hypotheses: *i)* ontology developers reuse hierarchical subtrees along with single terms, *ii)* the proportion of term pairs that have parent–child or sibling relations can be very high, especially if the reuse occurs from one main source ontology and *iii)* these terms are located on higher levels or upper-middle levels of ontological depth.

5.4. BioPortal Import Plugin Log Analysis

As was observed from our term reuse analysis across BioPortal ontologies, ontology developers only import terms from a small set of popular ontologies in BioPortal using the BioPortal Import Plugin. From our analysis of the logs, it is apparent that: *i)* ontology engineers have imported hierarchical subtrees of varying depths along with single terms, *ii)* the most common reuse structures are 2-level structures – parent–child structures (triangles with a higher opacity in Figure 8), and *iii)* these structures and terms are located in the higher and upper-middle levels of the ontological hierarchy.

Hence, we can say that the claims made from our clustering analysis (Section 5.3) are validated through our BioPortal Import Plugin log analysis. As future work, we plan to do a more formal validation of this finding. Moreover, for some ontologies that were common between both our analysis (e.g. NCIT, GO and FMA), we found a substantial similarity between some sub-clusters and the reuse structures extracted from the logs (results online). The similarity ranged between 70–100% for NCIT structures. This similarity can suggest either the ontologies developed using the BioPortal Import Plugin were saved back to BioPortal repository, or there are recurrence patterns in some ontologies that are reused frequently in different ontologies.

From this validation, we can postulate that our approach used for the two-phase clustering process, using the similarity equation and the term–term affinity matrices, accurately captures the thought process of the ontology engineer, when she reuses terms, and it can be coupled with the Bio-Portal Import Plugin to provide reuse recommendations in the future. The clustering only used the terms in the IRI reuse module, and might be biased towards OBO Foundry ontologies, and not generate enough UMLS recommendations (as they are seldom reused using the same IRI). Hence, our initial term–ontology matrix and the similarity equation will need to be extended to deal with this bias.

5.5. Future Work

All ontology development methodologies encourage reuse with several advantages, such as cost reduction, quality control, semantic interoperability, EHR mining and query federation, cited in favor of reuse [10,11,12,20]. However, our extensive analysis suggests that ontology developers do intend to reuse terms, but often, they are not able to do so correctly. Converting the intent for reuse into actual reuse can help increase term reuse, and reduce term overlap (Section 4.2).

We plan to provide personalized reuse recommendations for ontology developers through a WebProtégé plugin (http://webprotege.stanford.edu) [44]. The plugin will use our term–term affinity matrix (Section 3.4) and an item–based collaborative filtering method [45] to generate personalized recommendations for ontology developers, based on their target ontology and the engineering task at hand. These recommendations will be provided through a visual recommendation plugin built inside WebProtégé, where ontology developers can drag and select their terms of interest for reuse. This plugin may also keep developers informed, when the representation of the term in the source ontology changes.

We believe that our Web application will allow ontology developers to search for similar terms in other ontologies, while our visualization of overlap and reuse dependencies may guide developers to reuse terms in their own ontology based on the structure of ontologies in related domains. Our composite mappings approach may serve as a complement to the existing BioPortal mappings, which are currently generated through naive string matching algorithms [46]. We also plan to develop a term–centric visualization that summarizes everything known about a particular term in Bio-Portal, and presents it to developers and domain experts through an interactive interface. Our hope is that this visualization will enable ontology developers to serendipitously discover and reuse existing knowledge.

6. Conclusion

We estimated the level of reuse and overlap in a corpus of 337 ontologies from the BioPortal repository. We developed novel methods for detecting reuse and overlap in biomedical ontologies. Our findings show a term overlap of approximately 25.31–30.18%, and term reuse of less than 9%. Most ontologies reuse less than 5% of their terms from a small set of popular ontologies, with terms from several ontologies never being reused. We found strong indications that users actually intended to reuse terms, but in many cases they used incorrect representations. We also identified common error patterns in term reuse. Our hope is that the results of this work may be used to develop better guidelines and tool support with the aim to enhance reuse, and minimize overlap among biomedical ontologies.

Acknowledgments

The authors acknowledge Manuel Salvadores for providing a triplestore dump of BioPortal ontologies, and other members of the Protégé Group and the National Center for Biomedical Ontology for their input. This work is supported in part by grants GM086587 and GM103316 from the US National Institutes of Health.

References

1. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of medical informatics. 2008:67. [PubMed: 18660879]
2. Rubin DL, et al. Biomedical ontologies: a functional perspective. Briefings in bioinformatics. 2008; 9(1):75–90. DOI: 10.1093/bib/bbm059 [PubMed: 18077472]
3. Sioutos N, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics. 2007; 40(1):30–43. DOI: 10.1016/j.jbi.2006.02.013 [PubMed: 16697710]
4. Ashburner M, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000; 25(1): 25–29. DOI: 10.1038/75556 [PubMed: 10802651]

5. Stearns, MQ., et al. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. SNOMED clinical terms: overview of the development process and project status; p. 662
6. Suárez-Figueroa, MC. PhD thesis. Informatica; 2010. NeOn Methodology for building ontology networks: specification, scheduling and reuse.
7. Simperl E. Reusing ontologies on the Semantic Web: A feasibility study. Data & Knowledge Engineering. 2009; 68(10):905–925. DOI: 10.1016/j.datak.2009.02.002
8. Corcho O, et al. Methodologies, tools and languages for building ontologies. Where is their meeting point? Data & knowledge engineering. 2003; 46(1):41–64. DOI: 10.1016/S0169-023X0200195-7
9. Alexander CY. Methods in biomedical ontology. Journal of biomedical informatics. 2006; 39(3): 252–266. DOI: 10.1016/j.jbi.2005.11.006 [PubMed: 16387553]
10. Tudorache, T., et al. Knowledge Engineering and Management by the Masses. Springer; 2010. Ontology development for the masses: creating ICD-11 in WebProtégé; p. 74-89.
11. Rodrigues JM, et al. Sharing ontology between ICD 11 and SNOMED CT will enable seamless re-use and semantic interoperability. Studies in health technology and informatics. 2012; 192:343–346. DOI: 10.3233/978-1-61499-289-9-343
12. Kamdar MR, et al. ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of biomedical informatics. 2014; 47:112–130. DOI: 10.1016/j.jbi.2013.10.001 [PubMed: 24135450]
13. Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 11; 2007; 25:1251–1255. DOI: 10.1038/nbt1346
14. OBOFoundry. Inter-ontology Links. 2011. http://goo.gl/OSrSjP accessed March 01, 2015
15. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004; 32(suppl 1):D267–D270. DOI: 10.1093/nar/gkh061 [PubMed: 14681409]
16. W3C. OWL 2 Web Ontology Language Document Overview. 2012. http://www.w3.org/TR/owl2-overview/ accessed March 01, 2015
17. Whetzel PL, et al. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research. 2011; 39(suppl 2):W541–W545. DOI: 10.1093/nar/gkr469 [PubMed: 21672956]
18. Noy NF, et al. Ontology development 101: A guide to creating your first ontology. 2001
19. Cristani M, et al. A survey on ontology creation methodologies. International Journal on Semantic Web and Information Systems. 2005; 1(2):49–69. DOI: 10.4018/jswis.2005040103
20. Bontas EP, et al. Case studies on ontology reuse. Proceedings of the International Conference on Knowledge Management. 2005; 74 DOI:10.1.1.88.2772.
21. d'Aquin, M., et al. Modular ontologies. Springer; 2009. Criteria and evaluation for ontology modularization techniques; p. 67-89.
22. Pathak J, et al. Survey of modular ontology techniques and their applications in the biomedical domain. Integrated computer-aided engineering. 2009; 16(3):225–242. DOI: 10.3233/ICA-2009-0315 [PubMed: 21686030]
23. Xiang Z, et al. OntoFox: web-based support for ontology reuse. BMC research notes. 2010; 3(1): 175.doi: 10.1186/1756-0500-3-175 [PubMed: 20569493]
24. Courtot M, et al. MIREOT: The minimum information to reference an external ontology term. Applied Ontology. 2011; 6(1):23–33. DOI: 10.3233/AO-2011-0087
25. Nair, J., et al. The BioPortal Import Plugin for Protégé. Proceedings of the 2nd International Conference on Biomedical Ontology; 2011. CEUR-WS
26. Nair, J. BioPortal Import Plugin. 2014. http://goo.gl/LL75TR accessed March 01, 2015
27. Noy NF, et al. Creating semantic web contents with Protégé-2000. IEEE intelligent systems. 2001; 16(2):60–71.
28. Hanna J, et al. Simplifying MIREOT: a MIREOT Protégé plugin. The Semantic Web– ISWC. 2012
29. Wächter, T., et al. Proceedings of SWAT4LS. ACM; 2011. DOG4DAG: semi-automated ontology generation in OBO-edit and Protégé; p. 119-120.
30. Garcia-Santa, N., et al. Protege LOV Plugin. 2015. http://goo.gl/9fmTf7 accessed March 05, 2015

31. Linked Open Vocabularies (LOV). http://lov.okfn.org/dataset/lov (accessed October 09, 2015)
32. Matentzoglu, N., et al. The Semantic Web–ISWC. Springer; 2013. A snapshot of the OWL Web; p. 331-346.2013
33. Poveda Villalón M, et al. The landscape of ontology reuse in linked data. Proceedings of Ontology Engineering in a Data-driven World, Informatica. 2012
34. Ghazvinian A, et al. How orthogonal are the OBO Foundry ontologies? Journal of Biomedical Semantics. 2011; 2(2):1.doi: 10.1186/2041-1480-2-S2-S2 [PubMed: 21569604]
35. Kamdar, MR., et al. Investigating Term Reuse and Overlap in Biomedical Ontologies. Proceedings of the 6th International Conference on Biomedical Ontology, ICBO; 2015. p. 27-30.CEUR-WS
36. Tordai A, et al. Lost in translation? empirical analysis of mapping compositions for large ontologies. 2010:13–24.
37. Aleksovski, Z., et al. Managing Knowledge in a World of Networks. Springer; 2006. Matching unstructured vocabularies using a background ontology; p. 182-197.
38. Huang, Z. Clustering large data sets with mixed numeric and categorical values. Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining,(PAKDD); Singapore. 1997. p. 21-34.DOI:10.1.1.94.9984
39. Ng AY, et al. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems. 2002; 2:849–856.
40. Hastings, J., et al. Representing mental functioning: Ontologies for mental health and disease. ICBO 2012: 3rd International Conference on Biomedical Ontology; 2012. CiteseerDOI: 10.1.1.308.236
41. Hasnain, A., et al. The Semantic Web–ISWC. Springer; 2014. Linked biomedical dataspace: lessons learned integrating data for drug discovery; p. 114-130.
42. Ferreira Cordeiro, RL., et al. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2011. Clustering very large multi-dimensional datasets with mapreduce; p. 690-698.
43. Kuramochi, M., et al. Frequent subgraph discovery. Proceedings of International Conference on Data Mining; 2001; 2001. p. 313-320.IEEE
44. Tudorache T, et al. Web-Protege: A Lightweight OWL Ontology Editor for the web. OWLED. 2008; 432
45. Sarwar, B., et al. Proceedings of the 10th international conference on World Wide Web. ACM; 2001. Item-based collaborative filtering recommendation algorithms; p. 285-295.
46. Ghazvinian, A., et al. AMIA Annual Symposium Proceedings. Vol. 2009. American Medical Informatics Association; 2009. Creating mappings for ontologies in biomedicine: simple methods work; p. 198

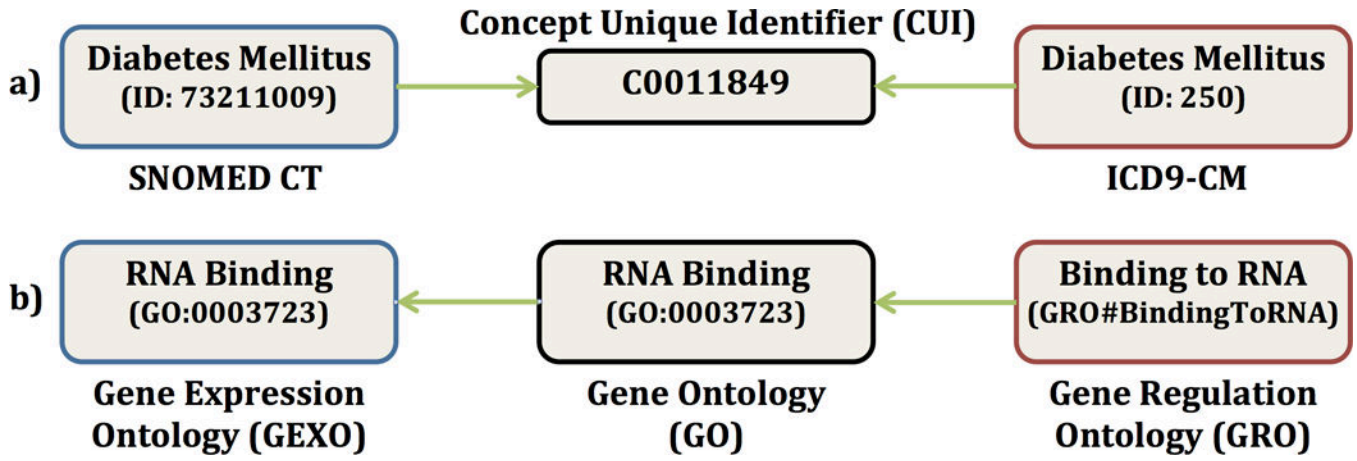

Fig. 1.

Types of Reuse: a) **CUI reuse**: *Diabetes Mellitus* terms in SNOMED CT and ICD-9CM are mapped to the same CUI, b) **IRI reuse**: *RNA Binding* defined in the GO ontology is reused in GEXO ontology using the same IRI; ***xref* reuse**: the latter term is reused in the GRO Ontology via a *xref* annotation

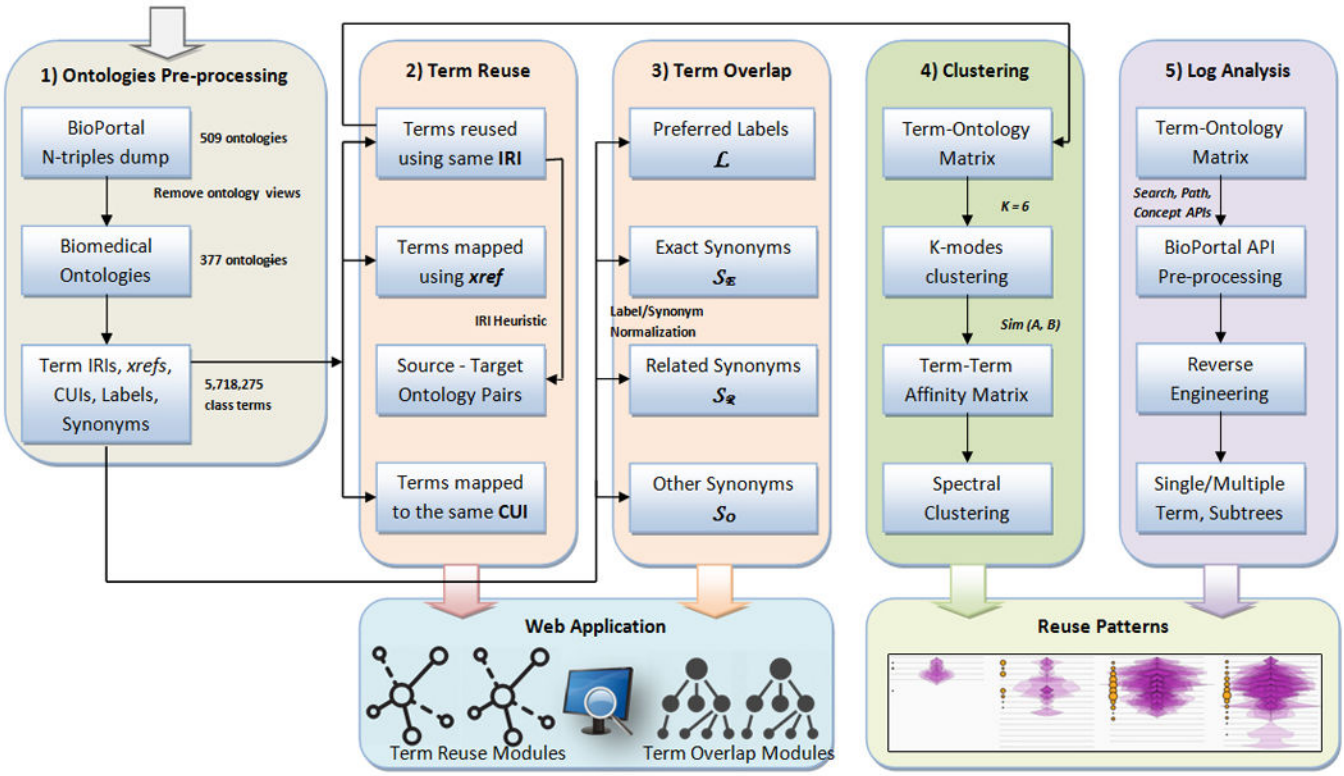

Fig. 2.

Worflow of all the steps required to estimate the average term reuse and overlap statistics across the BioPortal Ontologies, as well as clustering and BioPortal Import Plugin Log analysis to detect any reuse patterns. The steps of the workflow are: (1) Ontology Pre-processing, (2) Term Reuse, (3) Term Overlap, (4) Clustering, and (5) Log Analysis.

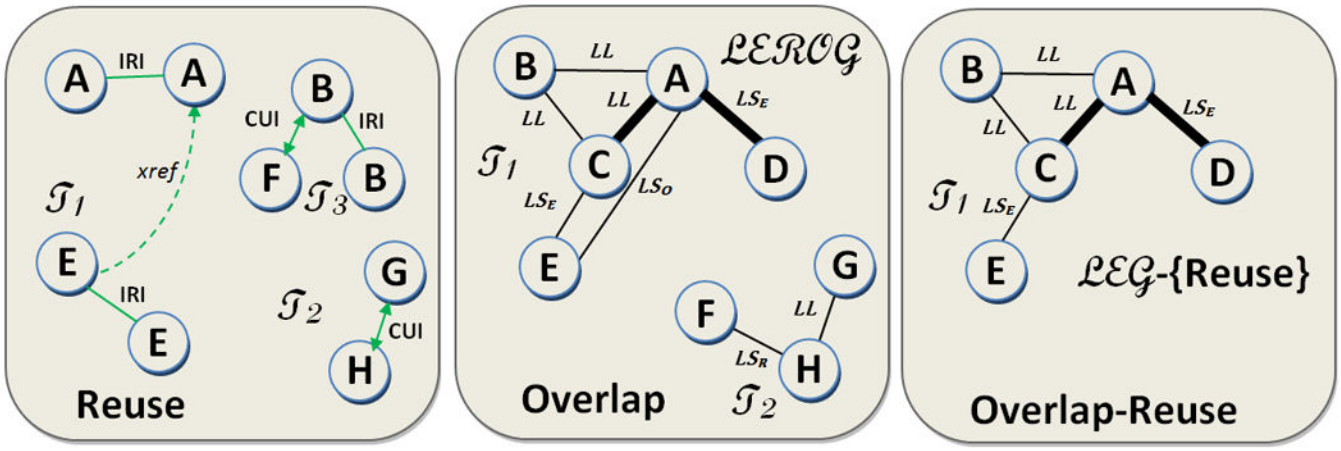

Fig. 3.

Cartoon representations of the **a)** *Reuse*, **b)** *Overlap* : $\mathcal{LEROG}$ and **c)** *Overlap – Reuse* : $\mathcal{LEG} - \{Reuse\}$ modules. In **a**) Terms A and E are defined in two ontologies using same IRI. The green, dotted arrow in *Reuse* module is a *xref* mapping from $E \rightarrow A$, whereas the green, bidirectional arrow means the terms G and H are mapped to same CUI. In **b**) and **c**) the two disjoint components $\mathcal{T}_1$ and $\mathcal{T}_2$ are composed of {*A,B,C,D,E*} and {*F,G,H*} terms respectively. The darkened path $C \rightarrow A \rightarrow D$ represents a sample composite mapping, formed by different edge types.

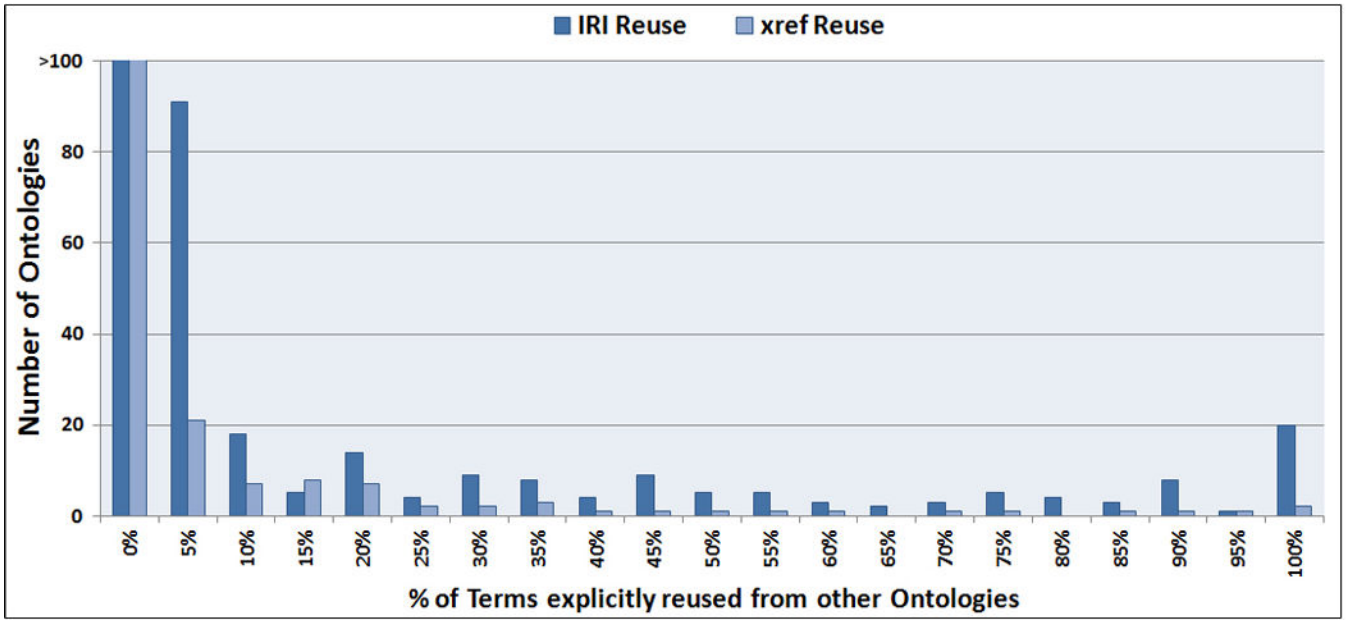

Fig. 4.

Histogram depicting the number of ontologies that reuse a given percentage (%) of terms **from** other ontologies in their current versions by the same IRI or *xref* annotation. Most ontologies reuse fewer than 5% of their terms.

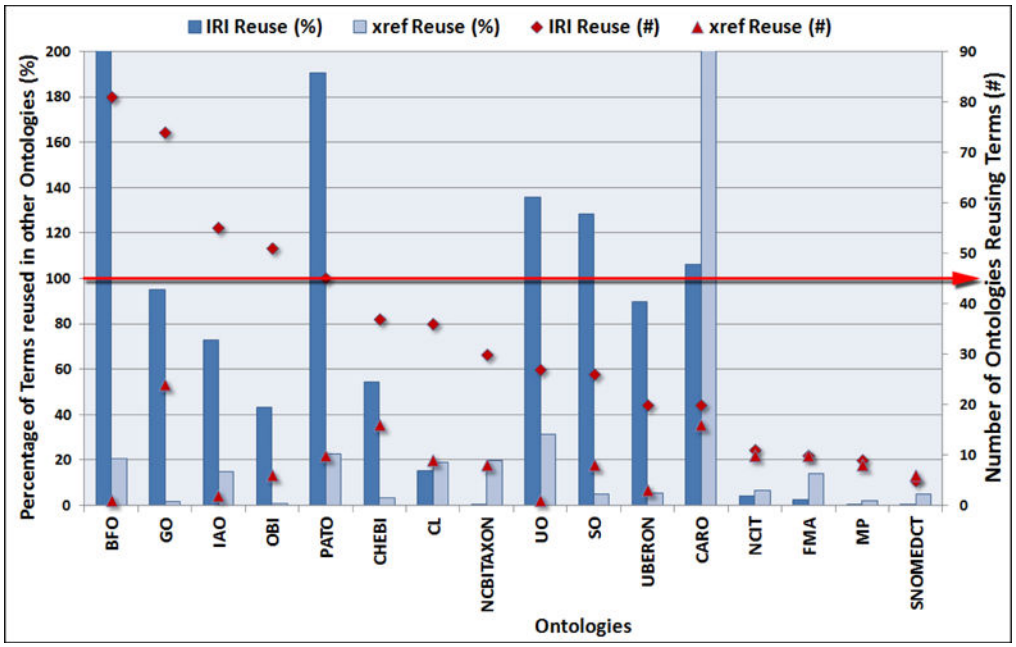

Fig. 5.

Top 16 ontologies whose terms are reused the most through *IRI* and *xref* constructs. Number of ontologies reusing (#) and percentage (%) of terms reused with respect to the terms in their current version.

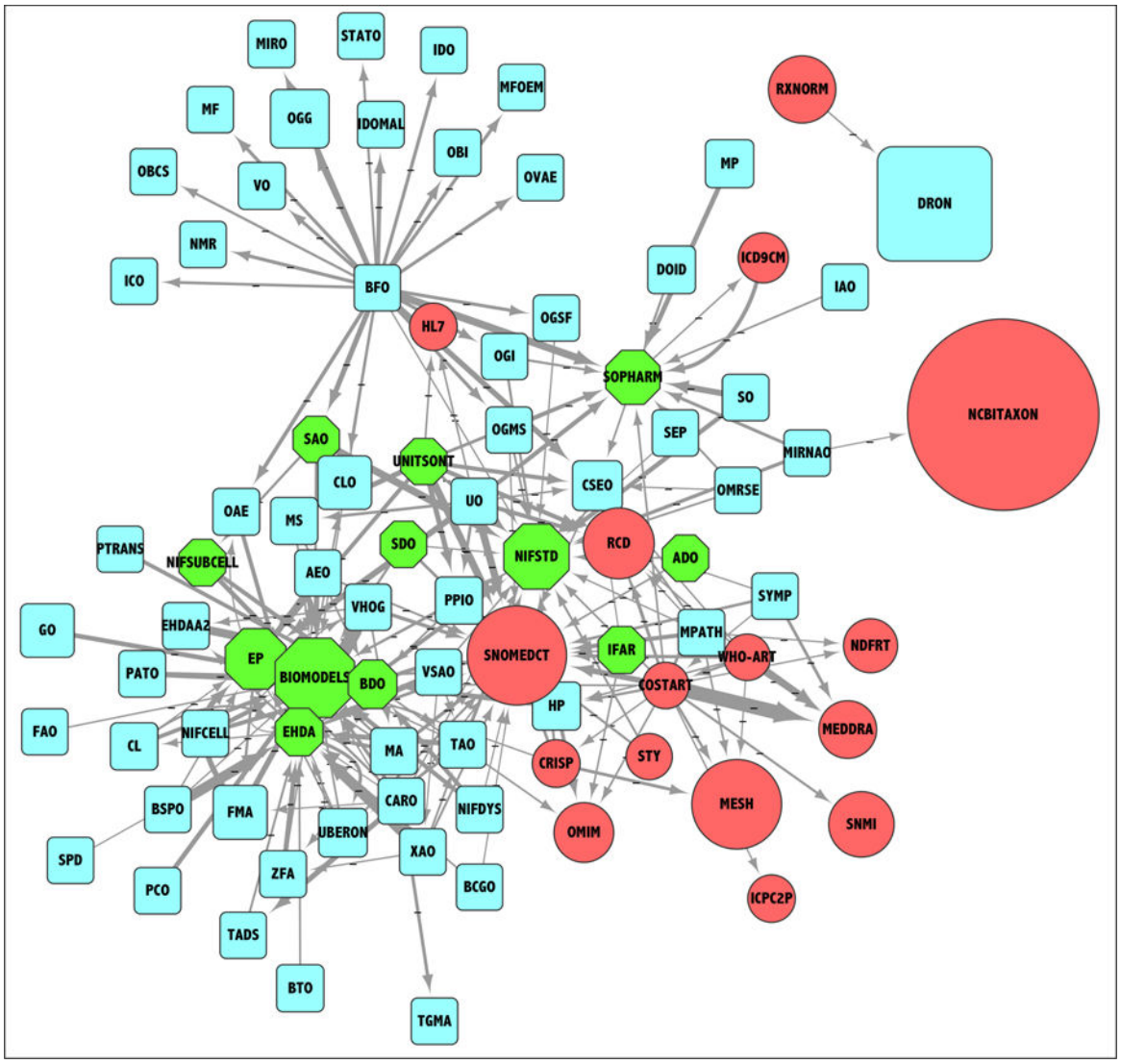

Fig. 6.

30% term overlap among different BioPortal ontologies. For simplicity, only the OBO Foundry member and candidate ontologies (blue squares), UMLS terminologies (red circles), and a few popular ontologies in BioPortal (green octagons) are shown here.

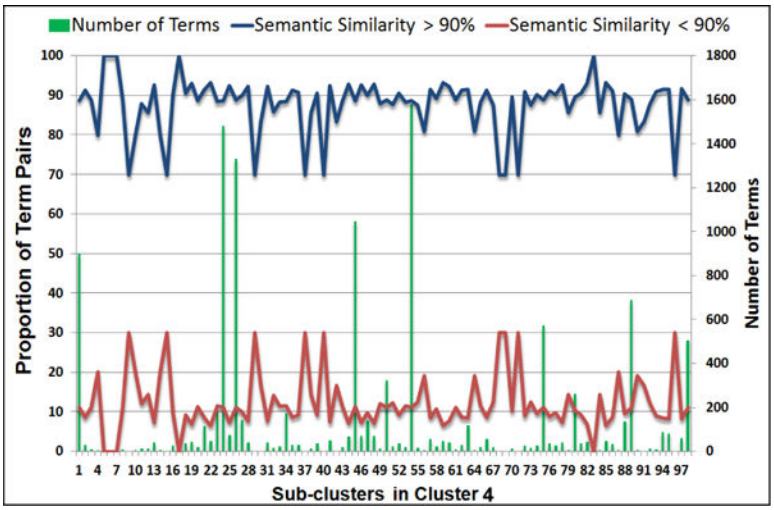

Fig. 7.
Proportion of term pairs with semantic similarity in a given range for each sub-cluster.

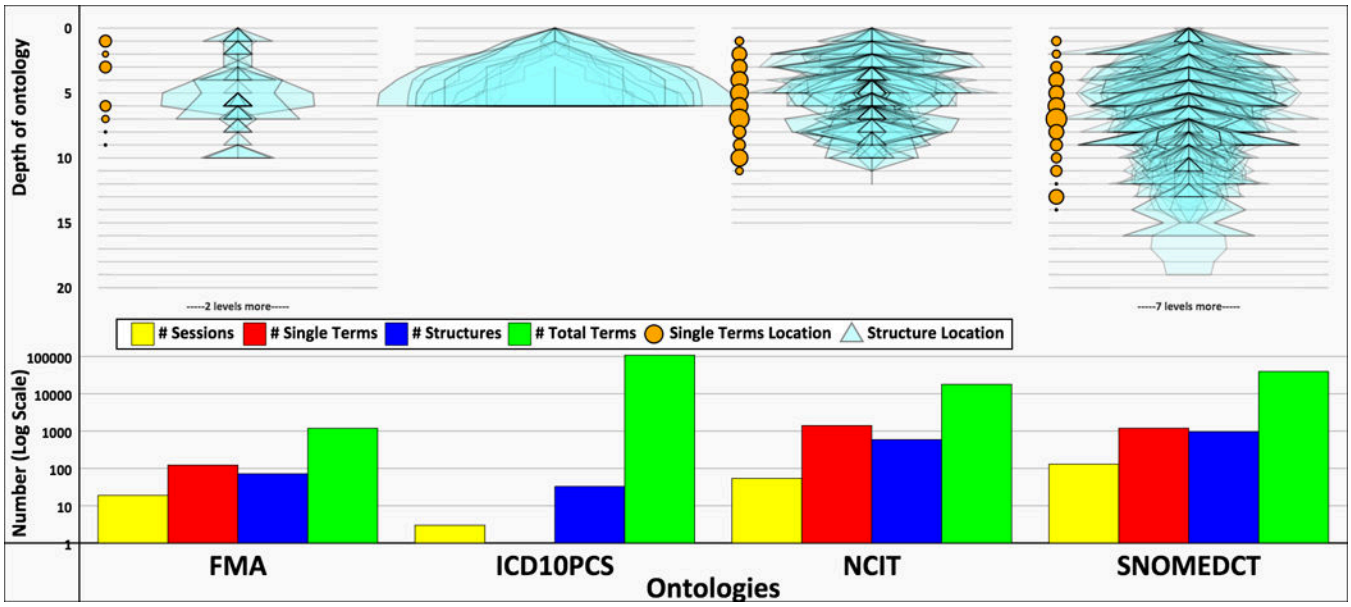

Fig. 8. BioPortal Import Plugin Log Analysis

Few ontologies that are reused the most through the BioPortal Import Plugin are shown — FMA, ICD10PCS, NCIT and SNOMED CT. The lower plot indicates the total number of sessions observed, the total number of single terms imported, the total number of structures imported, and the total number of terms imported in log scale. The upper plot indicates the content imported from each ontology spanning across its depth. Each structure imported is represented as a translucent polygon, whereas the single terms are grouped as circular shapes for each level.

Table 1

An example of a composite mapping. A column represents the term shown in the header. The content of a column contains different labels (preferred labels, synonyms, etc.) associated to the term. An arrow indicates that a label of a term is mapped to the label of another term. This example shows how we can map Term A defined in $\mathcal{O}_1$ to Term C defined in $\mathcal{O}_3$ using a composite mapping.

Term A ($\mathcal{O}_1$)	Term B ($\mathcal{O}_2$)	Term C ($\mathcal{O}_3$)
Heart Muscle →	Muscle of Heart	Myocardium
	Cardiac Muscle →	Cardiac Muscle

Table 2

Sources for labels and synonyms to generate composite mappings

Set	Source
$\mathscr{L}$	skos:prefLabel, rdfs:label, dc:title
$\mathscr{S}_{\mathscr{E}}$	OBO:hasExactSynonym, skos:altLabel
$\mathscr{S}_{\mathscr{R}}$	OBO:hasRelatedSynonym, OBO:IAO_0000118
$\mathscr{S}_{\mathscr{O}}$	OBO:hasNarrowSynonym, OBO:hasBroadSynonym, under IAO:000015, rdfs:comment, skos:definition

Table 3

An example of a hybrid component $\mathscr{T}_1$, composed of terms $\{t_i/i = 1, 2,\ldots, 7\}$. $\mathscr{T}_1$ can be broken into two smaller, relevant components $\mathscr{T}_{1a}$ and $\mathscr{T}_{1b}$ that are connected by an incorrect mapping caused due to a synonym of term t_3.

	Component ($\mathscr{T}_{1a}$)		Component ($\mathscr{T}_{1b}$)
t_1	Myocardium	t_4	Intercalated Disk
t_2	Cardiac Muscle	t_5	Intercalated-Disc
t_3	Heart Muscle	t_6	Discus Intercalatus
t_3	(Intercalated disc) →	t_7	Intercalated Disc

Table 4

Term overlap (actual and hybrid-adjusted) estimated for different overlap modules composed of different mappings.

Row #	Overlap Module	Terms #	Components #	Term Overlap (TO)	Hybrid Components # (Terms #)	Non-hybrid TO
1	$\mathscr{LG}$	2,230,636	781,007	25.39%	10 (1,119)	25.37%
2	$\mathscr{LEG}$	2,485,478	759,571	30.18%	1,187 (279,635)	25.31%
3	$\mathscr{LERG}$	2,565,928	755,816	31.65%	725 (361,120)	25.35%
4	$\mathscr{LER}\mathscr{S}\mathscr{G}$	2,475,905	744,314	30.28%	868 (289,090)	25.24%
5	$\mathscr{X}\mathscr{G}$	2,620,032	746,993	32.75%	270 (431,831)	25.21%
6	$\mathscr{LEG} - \{Reuse\}$	1,789,407	553,114	21.62%	182 (195,139)	18.21%
7	$\mathscr{LEG} - \{Reuse\ Intent\}$	1,232,149	284,499	16.57%	178 (192,475)	13.21%

Table 5

Primary ontological composition of the clusters

Cluster	Ontologies
Cluster 1	HINO, BIOMODELS, CHEBI, CCO, DRON, BDO
Cluster 2	GO, NIFSTD, GO-EXT, FYPO, CCO, NIGO, CL
Cluster 3	GWAS_EFO_SKOS, EFO, EFOGWAS, CCONT, CLO
Cluster 4	SYN, CSEO, SOPHARM, SNPO, IFAR, NCIT
Cluster 5	PHENOSCAPE-EXT, UBERON, NIFSTD, CL, CLO
Cluster 6	NIFSTD, ERO, DOID, CLO, NIFCELL, NIFDYS

Table 6

Different kinds of IRI representations observed in BioPortal ontologies and BioPortal Import Plugin logs.

Type	Source	Representation	Few Observed Examples
Versions	BFO	www.ifomis.org/bfo/1.1*	**(AERO)** Adverse Event Reporting Ontology
		www.ifomis.org/bfo/1.0	**(SAO)** Subcellular Anatomy Ontology
	NCIT	NCIT:C53037*	**(NCIT)** National Cancer Institute Thesaurus
		NCIT:Cerebral_Vein	**(CSEO)** Cigarette Smoke Exposure Ontology
Notations	FMA	OBO:FMA_31396*	**(VO)** Vaccine Ontology
		OBO:owlapi/fma#FMA_31396	**(BIOMODELS)** BioModels Ontology
		OBO:owl/FMA#FMA_31396	**(EP)** Cardiac Electrophysiology Ontology
		OBO:fma#Cartilage_of_inferior…	BioPortal Import Plugin Logs
Namespaces	BFO	www.ifomis.org/bfo/	**(ADO)** Alzheimer's Disease Ontology
		purl.obolibrary.org/obo/BFO_	**(IDO)** Infectious Disease Ontology
	SNOMED CT	ihtsdo.org/snomedct	**(SNOMED CF)** SNOMED Clinical Findings
		purl.bioontology.org/ontology/SNOMEDCT	**(IFAR)** Fanconi Anemia Ontology
	FMA	sig.uw.edu/fma#	**(BDO)** Bone Dysplasia Ontology
		purl.obolibrary.org/obo/FMA_	**(SDO)** Sleep Domain Ontology

* marks the recommended representation(s).

Listing 1

An anonymized excerpt of the BioPortal Import Plugin Logs

```
10. XX.XXX.XX - - [16/Dec/2011:14:26:12 -0800] “GET /bioportal/search/
Subthalamus/?ontologyids=1053& objecttypes = class & maxnumhits=20 HTTP/1.1”

10. XX.XXX.XX - - [16/Dec/2011:14:26:14 -0800] “GET /bioportal/path/44507/?
source=fma:Subthalamus&target=rootHTTP/1.1”

10. XX.XXX.XX - - [16/Dec/2011:14:26:14 -0800] “GET /bioportal/concepts/44507?
conceptid=http%3A%2F%2Fsig.uw.edu%2Ffma%23Anatomical_entityHTTP/1.1”
```