Analyzing user interactions with biomedical ontologies: A visual perspective

Maulik R. Kamdar, Simon Walk, Tania Tudorache and Mark A. Musen

Stanford Center for Biomedical Informatics Research, Stanford University, USA

Abstract

Biomedical ontologies are large: Several ontologies in the BioPortal repository contain thousands or even hundreds of thousands of entities. The development and maintenance of such large ontologies is difficult. To support ontology authors and repository developers in their work, it is crucial to improve our understanding of how these ontologies are explored, queried, reused, and used in downstream applications by biomedical researchers. We present an exploratory empirical analysis of user activities in the BioPortal ontology repository by analyzing BioPortal interaction logs across different access modes over several years. We investigate how users of BioPortal query and search for ontologies and their classes, how they explore the ontologies, and how they reuse classes from different ontologies. Additionally, through three real-world scenarios, we not only analyze the usage of ontologies for annotation tasks but also compare it to the browsing and querying behaviors of BioPortal users. For our investigation, we use several different visualization techniques. To inspect large amounts of interaction, reuse, and real-world usage data at a glance, we make use of and extend PolygOnto, a visualization method that has been successfully used to analyze reuse of ontologies in previous work. Our results show that exploration, query, reuse, and actual usage behaviors rarely align, suggesting that different users tend to explore, query and use different parts of an ontology. Finally, we highlight and discuss differences and commonalities among users of BioPortal.

Keywords: data visualization, ontology reuse, ontology exploration, user behavior, knowledge exploration, log analysis

1. Exploring Empirical Usage of Ontologies

Biomedical researchers have adopted ontologies for use in various tasks, such as knowledge management, data annotation, data integration, data exchange, decision support, and reasoning [1, 2]. These ontologies, which often serve as standard vocabularies for a specific domain, are exhaustively large. For example, the Chemical Entities of Biological Interest (ChEBI) ontology [3] contains more than 50,000 and SNOMED CT more than 300,000 entities. To support the biomedical community in finding and using ontologies, the National Center for Biomedical Ontology (NCBO) has developed BioPortal [4],[1] an open online repository of biomedical ontologies and terminologies. BioPortal currently hosts more than 500 biomedical ontologies. Users access BioPortal frequently in their work. In the first half of 2016, more than 215,000 unique IP addresses accessed the BioPortal website and submitted more than 2.52 million requests for various ontologies and services [5].

BioPortal provides two modalities through which researchers can access the content of the ontologies: (1) an interactive website (referred to in the rest of the paper as WebUI) that enables biomedical researchers to explore ontologies using a class hierarchy visualization (see Figure 1) [6]; and (2) an application programming interface (API) that allows researchers to programmatically query the repository for specific ontologies and entities, which allows them to perform tasks such as search, mapping and annotation [7]. Figure 2 shows an example of a BioPortal API request.

In the remainder of the paper we will refer to users *exploring* ontologies when users browse the ontologies and their classes using the BioPortal WebUI (Figure 1). We will refer to users *query-*

[1] `http://bioportal.bioontology.org`

ing ontologies when the users make the BioPortal API requests (Figure 2), usually to retrieve content for their application. Finally, *reuse* will refer to the situation in which an ontology uses a class defined in another ontology (see Section 3.3).

Problem. Despite the success and widespread adoption of BioPortal for biomedical and Semantic Web research during the last 10 years, we do not have a clear understanding of how researchers use BioPortal to explore, query and use ontologies in their own projects. Remedying this missed opportunity by investigating empirical usage data from BioPortal will allow BioPortal developers to better target their efforts to meet the needs of biomedical researchers. Through such an investigation, biomedical researchers can identify frequently used classes in their ontologies, while ontology engineers could concentrate their efforts on improving the content of highly accessed classes in their ontologies. We strongly believe that insights gained by such an investigation can guide the development of new interactive visualization methods and efficient semantic resource search and exploration methods. It will also enable the development of methods to profile users based on their behavioral characteristics and provide targeted recommendations.

We attempt to answer the following research questions in our work:

1. **RQ1: Do BioPortal WebUI exploration and API querying inform reuse?** Are the classes that users explore and access more often through the WebUI and API also reused more often in other ontologies?

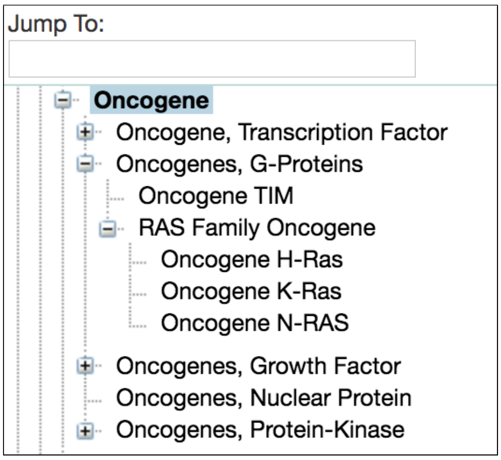
2. **RQ2: Do BioPortal WebUI exploration and API querying correlate?** Do users explore the same classes in the BioPortal WebUI that they query through the API?

3. **RQ3: Do BioPortal WebUI exploration and API querying inform usage?** Are the classes that users explore and access more often through the WebUI and API also used more in downstream applications?

The research questions were formulated, in part, due to our prior research, categorizing exploration behaviors from BioPortal WebUI logs and analyzing ontology reuse across biomedical ontologies [5, 8].

Approach. In this paper, we present an empirical investigation to help improve our understanding of how researchers *i)* explore, *ii)* query, and *iii)* reuse biomedical ontologies and terminologies from BioPortal [5, 8, 9]. We hypothesize that users of BioPortal mainly use the WebUI to explore classes of ontologies to determine their utility for a specific task before programmatically querying (API) and using these classes for ontology reuse or downstream applications. For example, a user might first explore the `Oncogene` class of the NCI Thesaurus [10] in the BioPortal WebUI (Figure 1), and then, if she deemed it appropriate for her application, she would access the `Oncogene` class through the BioPortal API. Specifically, we apply several visualization techniques to analyze commonalities and differences among information consumption strategies across different interfaces (i.e., WebUI and API) and purposes (i.e., exploration, querying, reuse and downstream applications).

For each ontology, we analyze correlation statistics between WebUI exploration data, API querying data and reuse data. To inspect and compare large amounts of exploration, query and reuse data, we make use of and extend PolygOnto [8], a visualization method in which an ontology is represented as an abstract geometrical polygon. The PolygOnto visualization method enables a quick comparison of individual usage and information consumption



Jump To:

- Oncogene
  - Oncogene, Transcription Factor
  - Oncogenes, G-Proteins
    - Oncogene TIM
    - RAS Family Oncogene
      - Oncogene H-Ras
      - Oncogene K-Ras
      - Oncogene N-RAS
  - Oncogenes, Growth Factor
  - Oncogenes, Nuclear Protein
  - Oncogenes, Protein-Kinase

Figure 1: The Oncogene subtree of the National Cancer Institute Thesaurus (NCIt) displayed in class hierarchy of the BioPortal WebUI. The "Jump To" field allows users to quickly select a class in the hierarchy.

http://data.bioontology.org/search?q=Oncogene &ontologies=NCIT

Figure 2: An example BioPortal API request to retrieve all classes called "Oncogene" from the NCI Thesaurus.

strategies across different interfaces and ontologies. Through three scenarios, we further analyze whether user interactions in BioPortal inform real-world applications of ontologies.

We discuss some of the key findings of our empirical analysis with respect to ontology users' exploration and querying behavior, and highlight opportunities for further investigations. All results presented in this paper, including PolygOnto and other visualizations we have developed for each ontology, are available online at:
http://onto-apps.stanford.edu/vision.

The datasets used in this study have also been published as TSV and RDF files at:
http://onto-apps.stanford.edu/bionic under the Creative Commons CC-BY-NC-SA license [11].

The remainder of this paper is organized as follows: Section 2 describes work related to the analysis of user logs in the context of the Semantic Web, and to ontology visualization. In Section 3, we characterize the datasets used in the different analyses. In Section 4 we outline the methods used to compute and visualize correlation statistics, and delve deeper into the PolygOnto visualization method. We present the results of our empirical investigation in Section 5. Finally, in Section 6, we discuss few key findings regarding the ways in which users explore and query biomedical ontologies.

2. Related Work

2.1. Log Analysis to Characterize User Behavior

Several studies have tried to identify ways in which users interact with ontologies and ontology editors in the context of collaborative ontology development. These studies have used the data provided by logs of user activity in collaborative ontology development tools. Strohmaier et al. [12] conducted an empirical investigation using user activity logs to measure the impact of collaboration on ontology-engineering projects. The authors developed several new metrics to quantify different aspects of the hidden social dynamics that take place in these collaborative ontology-engineering projects from the biomedical domain. Falconer et al. [13] investigated and classified users according to different roles in collaborative projects by clustering users according to the types of changes they contributed.

Debruyne et al. [14] used different characteristics (reputation sensors) to measure the reputation of users in collaborative ontology-engineering projects with the goal of identifying "leaders" that drive activity, quality or social interactions. Using a combination of k-means and the GOSPL methodology [14], Van Laere et al. [15, 16] classified users by clustering interactions that users engage in while engineering an ontology. Vigo et al. [17] analyzed eye-tracking data and event logs from the Protégé ontology editor to identify common user activity patterns, and proposed guidelines for bulk editing and modifications to the Protégé user interface.

In 2013, Wang et al. [18] applied association-rule mining to the change-logs of several different collaborative ontology-engineering projects to extract edit patterns, which were then used to predict the next change actions in the corresponding projects. Similarly, Walk et al. [19, 20, 21] used (higher-order) Markov chains to study user-editing trails of ontology-engineering projects to predict the action a user is most likely to conduct next. Pesquita et al. [22] leveraged the location and specific structural features of edit trails to show that these features can be used to determine where the next change is going to take place in the Gene Ontology.

Recently, Walk et al. [5] conducted a first empirical study to cluster and analyze how users browse ontologies in NCBO's BioPortal [4]. They discovered a total of 7 different exploration strategies and showcased how certain attributes of an ontology can influence user interactions.

Yet, we still do not fully understand how users interact with ontologies across different modes of access and purposes. Moreover, we lack tools that allow ontology engineers, researchers and ontology repository maintainers to visualize and investigate these interactions between users and ontologies.

2.2. Ontology Visualization Methods

Over the years, researchers have proposed several methods for visualizing ontologies. Some of these methods are informed by the different requirements of end users, and also by the size and the complexity of the ontology [23]. Two groups of methods are widespread for ontology visualization: Indented Tree (e.g., as used in BioPortal, or the Protégé ontology editor [24]) and Graph-based methods (e.g., as used in VOWL [25]). Indented tree visualizations are more organized and familiar to novice users. At the same time, graph-based visualizations are more intuitive without visual redundancy, particularly for ontologies with multiple inheritance [6]. Kamdar et al. [26] have used graph-based visualizations of domain-specific ontologies to help do-

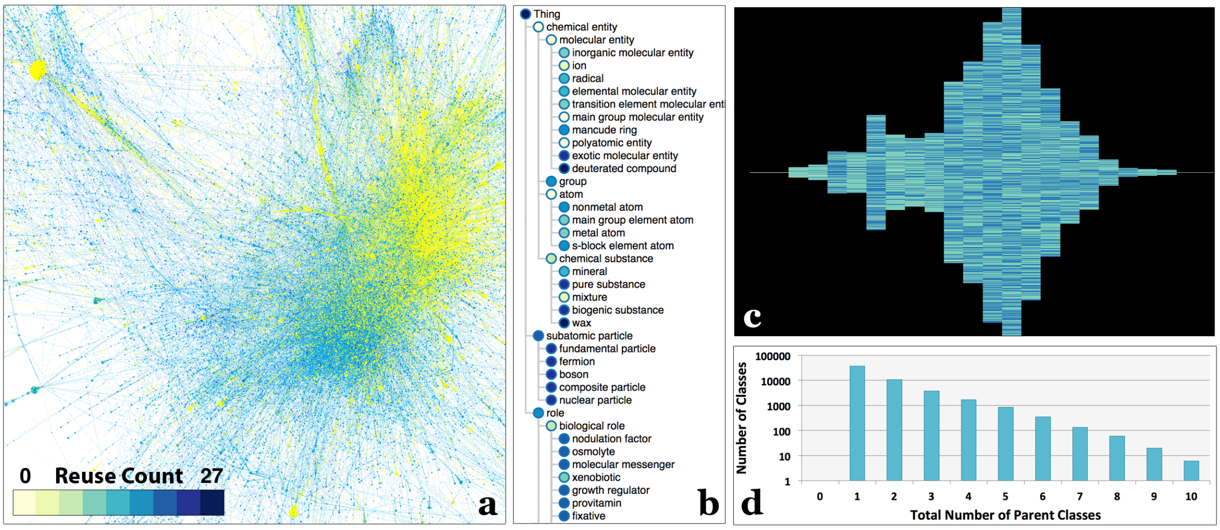


Figure 3: **We visualize the January 2015 version of ChEBI ontology using 3 methods. a)** Graph-based visualization, where nodes represent classes and edges represent *subClassOf* relations. The nodes are colored based on the number of ontologies reusing the underlying class. **b)** Indented Tree visualization. **c)** Implicit Hierarchy visualization, where classes are arranged horizontally based on maximum path distance from the root class and vertically based on proximity to the parent classes in the previous layer. **d)** Histogram representing the distribution of total number of parent classes for each class in ChEBI.

main experts formulate queries, especially for cases in which the experts were not familiar with the complex syntax and semantics of query languages.

Baehrecke et al. [27] have also experimented with other methods, such as 3D visualizations (e.g., Ontosphere [28]) and implicit hierarchy visualizations [29] (e.g., TreeMaps [30]) to visualize large ontologies. Implicit hierarchy visualizations do not use traditional 'node-link'-based approaches to represent hierarchy. They also rely on other attributes such as node sizes and positions. However, these methods may not be very efficient to provide a holistic view of an ontology, to visualize ontologies with multiple inheritance, or to visualize count statistics (e.g., number of ontologies reusing a class).

To demonstrate the different types of visualization techniques and discuss some of their limitations, we applied three different visualization methods to the ChEBI ontology (January 2015 version): Graph (Figure 3a), Indented Tree (Figure 3b), and Implicit Hierarchy (Figure 3c). We color-code the class nodes with respect to the number of times a class is reused in other ontologies using the same Internationalized Resource Identifier (IRI) [8]. The color scale is indicated in Figure 3a. As the underlying ontology had 54,847 classes, the Graph and Indented Tree visualizations become hard for users to explore without additional tools (such as zoom and drag) and impose a higher cognitive load on the users. The Implicit Hierarchy visualization organizes classes horizontally in different layers based on the maximum path distance from the root class by traversing the *subClassOf* links, and vertically such that classes in a lower layer are placed close to the position of their parents in the previous layer. However, as Figure 3d shows, ChEBI and other biomedical ontologies may have classes with several parents. Moreover, such plots can only show count statistics and not patterns (such as a group of classes reused together).

The Implicit Hierarchy visualization method can display the hierarchical structure of the ontology. The shape of the visualization bears resemblance to Violin plots [31] or Bean plots [32]. Such plots can effectively visualize the density distributions of data along with summary statistics and individual data points. They can aid in revealing the structure of the dataset, discovering anomalies in a dataset (e.g., bimodal distributions) and in the comparison of different datasets. Instead of visualizing each class in an ontology and its corresponding count in the dataset, as in the current Implicit Hierarchy visualization methods, a higher frame of abstraction may be useful to discover patterns in our data.

Recently, pictograph visualizations of large datasets have been demonstrated to make it easier for the user to remember the data and to improve discovery of relevant information, compared with minimalist charts [33]. The Electronic Fluorescent Pictograph visualization method [34] paints gene expression data from large-scale microarray datasets onto pictographic representations of the experimental samples used to generate the datasets.

The PolygOnto visualization method [8], which we extended and used for this study, is inspired by the Implicit Hierarchy and pictograph visualization methods. In essence, the PolygOnto method generates a pictorial representation of the ontology that can be used both to compare hierarchical structures of ontologies as well as to efficiently visualize exploration and query patterns for several users across large ontologies. In a previous study [8], we used a version of this method to visualize reuse patterns extracted from BioPortal Import Plugin logs [7].

3. Datasets

For this study, we use four different sources of data: *i)* a dump of BioPortal ontologies, *ii)* the reuse statistics generated by Kamdar et al. [9, 8], *iii)* the WebUI and API requests, which are logged when users explore and query BioPortal, and, *iv)* ontology-based data annotations in the NHGRI GWAS Catalog [35] and the Life Sciences Linked Open Data cloud [36, 26].

3.1. Ontology Datasets

We obtained a triplestore dump of the BioPortal ontologies in N-triples format that contained 509 distinct ontologies as of January 1, 2015. We extracted statistics for WebUI clicks and API query data from the BioPortal access logs for the period January 2013–June 2016.

For each class in each ontology, we assigned an attribute *maximum depth* as the maximum path distance from the given class to the root class (`owl:Thing` in OWL ontologies) by traversing the *subClassOf* links. For example, the maximum depth for classes in the example ontology depicted in Figure 4a is 5.

We removed ontological views (i.e., $\mathcal{O}_1 \subseteq \mathcal{O}_2$), as well as ontologies whose classes are never reused by other ontologies, and ontologies that do not reuse classes from other ontologies (cf. Section 3.3). We also exclude those ontologies that had less than 10 unique users that explored or queried them using the BioPortal WebUI and API respectively. After applying these filtering conditions, we were left with 115 distinct biomedical ontologies: 48 OWL ontologies, 8 OBO Foundry [37] member ontologies (e.g., GO, CHEBI), 38 OBO Foundry candidate ontologies (e.g., OGMS, HP) and 21 UMLS [36] Terminologies (e.g., SNOMED CT, ICD-9).

Table 1: Characteristics of the BioPortal access logs.

Feature	**WebUI**	**API**
Unfiltered Requests (ca.)	18.9M	346.2M
Class Requests (ca.)	5.4M	67.2M
Unique IPs	1,030,746	205,809
Observation periods (ca.)	3.5 years	3.5 years

3.2. WebUI Exploration & API Query Logs

For the analyses presented in this paper, we collected a total of 18.9M WebUI requests and 346.2M API queries from the BioPortal Apache access logs between January 2013 and June 2016 (Table 1). From each request, we used the information about when it was received by the server (timestamp), who submitted the request (IP address), and which resource was requested (the URL). Similarly to Walk et al. [5], we filter and blacklist IPs that belong to search-engine bots, crawlers and spiders (e.g., GoogleBot or Yahoo! Slurp) and remove all requests that do not contain a valid ontology abbreviation and class ID. This reduces the dataset to 5.4M WebUI requests and 67.2M API queries.

For our analyses, we define a *user* to be a unique, distinct IP address, that sends a request using either the BioPortal WebUI or the BioPortal API. Note that this assumption has a fallacy, as different users may share a common IP address or the same user may be assigned multiple IP addresses over consecutive visits of BioPortal. As we are mainly interested in the comparison of usages and interactions through the different modes (i.e., WebUI and API), the impact of this limitation is negligible.

3.3. Reuse Logs

Previously, Kamdar et al. computed reuse statistics across all biomedical ontologies in BioPortal [8, 9]. In this research, we use two different reuse constructs that are commonly found across biomedical ontologies: *i) IRI* - two classes share the same Internationalized Resource Identifier, and *ii) CUI* - two classes are mapped to the same UMLS Concept Unique Identifier. The reuse measures were computed for each class in every ontology in BioPortal.

From the previously computed reuse measures, we used the following in our analyses:

1. The number of ontologies reusing a class via the same IRI.
2. The number of ontologies reusing a class with the same CUI mapping.
3. The sets of classes from one ontology reused in other ontologies using the same IRI.
4. The sets of classes from one ontology whose CUI are mapped to classes in other ontologies.

3.4. Ontology-Annotated Datasets

For research question **RQ3: Do BioPortal WebUI exploration and API querying inform usage?**, we present real-world research scenarios in the biomedical domain that use biomedical ontologies to annotate datasets for knowledge management, data integration and search. We use our analyses and visualization methods (Section 4.1 & 4.3) to compare and correlate the use of the Experimental Factor Ontology (EFO) and the Chemical Entities of Biological Interest (ChEBI) ontology with the BioPortal WebUI exploration and API querying user behavior.

3.4.1. Catalog of Genome-wide Association Studies

The NHGRI GWAS Catalog is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies (GWAS). The studies and the associated traits are annotated with Experimental Factor Ontology (EFO) classes (e.g., lung adenocarcinoma – `EFO:0000571`) for enhanced search and interoperability. We downloaded the entire NHGRI GWAS Catalog and grouped EFO classes used to annotate the studies. For this research, our dataset includes 3,417 GWAS entries with 4,542 annotations and 1,375 distinct EFO classes.

3.4.2. Life Sciences Linked Open Data Cloud

Using Semantic Web technologies and linked data principles, biomedical researchers have published biomedical data sources as RDF graphs to create the Life Sciences Linked Open Data (LSLOD) cloud [36, 26]. For this research, we have identified 12 different LSLOD sources, such as PubChem [38], DrugBank [39], and PharmGKB [40], that have been used for various biomedical research tasks. For knowledge management and data integration, these sources annotate or cross-reference their schema elements and individual instances with classes from the Chemical Entities of Biological Interest (ChEBI) ontology. We retrieved and normalized 134,063 ChEBI-based annotations from these 12 sources using a distributed querying framework over the LSLOD cloud [41].

3.4.3. PubChem Database of Biological Assays

The PubChem database uses classes from the ChEBI ontology to annotate compounds that are used in biological assays [38]. We query the PubChem database to retrieve count data on the number of assays that use a particular compound. We retrieved 1,387,027 PubChem experiments (unique assay–compound pairs) with 44,305 unique ChEBI-annotated compounds.

4. Methods

4.1. Computing Exploration, Query and Reuse Statistics

We generated a graph structure for each ontology using the *subClassOf* relations and traversed this graph to determine the longest paths for each class from the root (generally, `owl:Thing`) in order to determine their maximum depth attribute.

For every class in the 115 ontologies in our dataset (Section 3.1), we extracted the following 4 attributes from the reuse, WebUI and API request logs: *i)* IRI reuse, *ii)* CUI reuse, *iii)* Unique IP requests using WebUI and *iv)* Unique IP requests using API. Note that OWL and OBO Foundry ontologies contain only IRI reuse, whereas UMLS terminologies contain only CUI reuse [8]. Therefore, we computed a single attribute *reuse* that reflects the IRI reuse in OBO and OWL ontologies, and CUI reuse for UMLS ontologies.

For each ontology, we computed three proportions of classes: (1) the proportion of classes accessed using the WebUI; (2) the proportion of classes accessed via the API; and (3) the proportion of classes that were reused in other ontologies. To calculate the proportions, we used the following generic formula:

$$Proportion_C(\mathcal{O}) = \frac{\sum_{t=1}^{N} \mathbb{1}[C(t) \geq 1]}{N} \qquad (1)$$

In Equation (1), N indicates the total number of classes in ontology $\mathcal{O}$. C indicates the count statistic used (i.e., $C(t)$ indicates the number of times a class t is either accessed or reused).

To evaluate our hypothesis that biomedical ontology developers tend to reuse the same classes that biomedical researchers explore via the WebUI and query via the API, we compute three pairs of Spearman correlation and Jaccard similarity statistics across all ontologies using the following comparisons: *i)* Reuse versus WebUI Access, *ii)* Reuse versus API Access, and *iii)* WebUI Access versus API Access. The first two comparisons indicate whether the classes reused in other ontologies are similar and correlate with the way users explore and query ontologies. The third comparison provides initial insights into the differences between information consumption strategies using the BioPortal WebUI and the API.

$$S_A = \{t | C_A(t) \geq 1\} \; ; \; S_B = \{t | C_B(t) \geq 1\} \qquad (2)$$

$$Jaccard_{AB}(\mathcal{O}) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \qquad (3)$$

The Jaccard similarity statistic for a particular comparison between any two attributes A and B (i.e. Reuse, WebUI Access or API Access; Equation (2)) can be computed using Equation (3). For the attributes being compared, we determine two sets of classes S_A and S_B, such that the count statistic for each class for the corresponding attribute (e.g. $C_A(t)$) must be greater than or equal to 1. The exact count data is not used here (e.g. the number of requests for a particular class).

$$Spearman_{AB}(\mathcal{O}) = \frac{\mathtt{cov}(rC_A, rC_B)}{\sigma_{rC_A}\sigma_{rC_B}} \qquad (4)$$

For computing the Spearman correlation statistic between any two attributes A and B, we use the count data to generate ordered rankings of the classes for each attribute (e.g. rC_A). The Spearman correlation statistic is calculated using Equation (4), where $\mathtt{cov}(rC_A, rC_B)$ denotes covariance of the ordered rankings, and σ_{rC_A} denotes standard deviation. Spearman correlation helps to minimize the impact of extreme outliers through rankings.

4.2. Comparing Ontology Usage in Annotations

We use the ontology-annotated datasets (Section 3.4) to generate additional real-world application attributes for ChEBI and EFO.

For each class in the ChEBI ontology, we have the following usage attributes from the LSLOD sources: *i)* number of sources that use the class for annotation or cross-reference, and *ii)* number of PubChem assays that experiment with a given compound cross-referenced to the class. Similarly, for each class in the EFO ontology, we have the number of studies in the NHGRI GWAS Catalog that are annotated with the class.

To determine if user behavior and ontology reuse inform ontology usage in the downstream applications, we compute the Spearman correlation and Jaccard similarity statistics between the real-world usage attributes, and the access and reuse attributes (Section 4.1). However, not all classes in a given ontology may be relevant for the goals of the downstream application (e.g. EFO ontology includes anatomy classes that are not relevant in GWAS studies). Hence, we also compute an "Adjusted Spearman Correlation" statistic that limits the set of attribute classes in the analysis to only those present in our annotated datasets.

4.3. Visualizing Ontologies using PolygOnto

In this paper we make use of and extend the PolygOnto visualization technique—which we initially developed to study reuse in biomedical ontologies [8]—to visually inspect exploration, query, reuse and usage data of biomedical ontologies on BioPortal. In PolygOnto, an ontology is represented as an abstract geometrical polygon that displays count-based attributes of the classes while at the same time retaining the hierarchical structure of the ontology. For example, a PolygOnto visualization for reuse would use the reuse count for a class, and show the class in the context of the class hierarchy. Figure 7 shows several examples of the PolygOnto visualization.

The main idea behind PolygOnto is that rather than individually representing each class in the ontology and its corresponding count-based attributes (cf. Section 4.1), we perform an abstraction based on the maximum depth attribute of the classes. Specifically, we aggregate different classes that are at the same maximum depth from the root class in the ontology. We then symmetrically align each depth layer. The breadth of each layer is proportional to the number of classes in the layer.

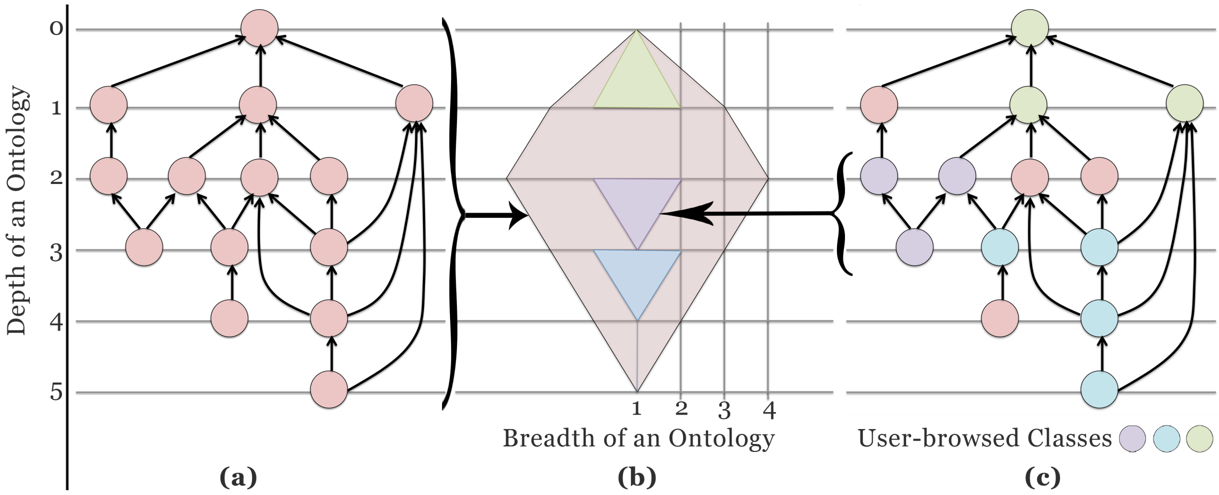


Figure 4: **Converting the ontological hierarchy to a PolygOnto visualization. a)** The directed acyclic graph structure of an example ontology is shown in the first figure, using a node-link diagram, where the nodes represent the classes in the ontology and the connecting links represent the *subClassOf* relations between the classes. **b)** The directed acyclic graph is converted to an abstract glyph by considering the maximum depth of any class from the root class (e.g., `owl:Thing`). Different classes are aggregated based on this maximum depth (e.g., 4 classes at depth 2) and are aligned symmetrically at each depth layer, where the breadth is proportional to the number of classes in the layer. **c)** Finally, user behaviors are also converted to such glyphs and overlaid upon the ontology glyph, to generate a PolygOnto. In this figure, 3 users (characterized by different color nodes) explore different classes of the ontology to generate the colored pictographs in **b)**. Generally, triangles indicate "parent–children" substructure and inverted triangles indicate "child–parents" substructure, with some exceptions.

Figure 4 shows the process of converting an ontology to a PolygOnto visualization. We convert the directed acyclic graph structure of an ontology to a polygonal glyph (i.e., a geometric structure), as shown in Figure 4 (a,b). Hence, the height of the glyph indicates the number of hierarchical layers in the ontology and the width of the glyph indicates the breadth of the ontology at each hierarchical layer. The edge, which connects two adjacent vertices in the polygonal glyph, indicates the increase or decrease of class count between layers. Classes in subsequent layers always have parent–child (or *subClassOf*) relations. For better proportionality, we consider the breadth of the layer to be log-scaled in the number of classes at that layer.

$$Breadth_i = \begin{cases} 2 * \gamma * log_e(\frac{n_i - 1}{2} + e) & \text{if } (n_i > 1) \\ 0 & \text{if } (n_i = 1) \end{cases} \tag{5}$$

For n_i classes in the i^{th} layer of the polygon, the breadth is calculated as shown in Equation (5). For visual simplicity, layers with single classes culminate as a point (i.e., $breadth = 0$). γ relates to the minimum breadth of the polygon at each layer. That is, a hierarchical layer with 2 classes has a breadth of $2.35 * \gamma$.

We use PolygOnto to visualize user exploration, querying or reuse patterns over the polygonal glyph of the ontology. Figure 4c depicts how we convert a user exploration pattern to a PolygOnto visualization. Suppose three different users explore three different sets of classes in an ontology. The nodes explored by a user are shown in the same color. The exploration paths are converted to polygonal glyphs using a similar method as described before. For example, the exploration behavior of the purple user includes a class (depth 3) and its two parent classes (depth 2), and is represented as an inverted triangle glyph at depth 2 (as the parent classes have a maximum depth of 2). Similarly, the exploration behavior of the green user includes the root class and its two child classes and is represented as a green triangle glyph at depth 0.

While it may seem intuitive that triangles represent some form of parent–child structure, it is worth noting that this may not be true for all cases (e.g., blue user). Straight lines along the vertical axis may represent a single parent–child pair. Note that we only consider classes explored together (co-occurrence) by the same IP address and neglect temporal and sequential information. Each glyph in PolygOnto has an opacity value smaller than 1, so that the underlying glyphs in the visualization can be seen. Hence, an opaque glyph at a particular layer indicates that the corresponding pattern (but not necessarily the same group of classes) was observed several times. The processing and visualization of other count-based attributes for classes, for example the number of queries via the API or how often a class was reused, follows analogously.

In Equation (5), we chose the natural logarithm over a base 10 logarithm because coefficients on the natural-log scale are directly interpretable as approximate proportional differences [42]. That is, minute differences in the breadth of the overlying pattern glyphs will be visible in PolygOnto visualization. We sum the Euler's constant e, to ensure that the breadth of the layer is a positive number. The parameter γ can be adjusted by the user—an optimal value will allow the overlying pattern glyphs (with different class counts in a layer) to be visually discernible, while not being broad.

5. Results

We demonstrate some of the key findings from our empirical analysis of the exploring, query, usage and reuse behavior of users by visually inspecting large interaction log data.

5.1. *Exploration & Query Statistics across biomedical Ontologies*

For each class in an ontology, we determined the unique number of IP requests made either using the BioPortal WebUI or API during the period of 2013–2016, and computed the proportion of classes that were accessed at least once during this time period (see Section 4.1).

The main scatter plot in Figure 5a visualizes the proportion of classes accessed using the BioPortal WebUI ($Proportion_{UI}(\mathcal{O})$) versus the proportion of classes accessed using the BioPortal API ($Proportion_{API}(\mathcal{O})$). Each ontology is visualized as a node in the scatter plot. The shape of each node depends on the type of ontology—OBO (square), UMLS (circle) and OWL (diamond). The size of the each node is proportional to the number of classes in ontology. The Spearman Correlation statistic for the *WebUI Access versus API Access* comparison is mapped to a yellow–blue color scale.

The horizontal scatter plot in Figure 5b aligns with Figure 5a along the $Proportion_{UI}(\mathcal{O})$ x-axis, visualizing the number of unique users that access a given ontology through the BioPortal WebUI. Similarly, the vertical scatter plot in Figure 5c aligns with Figure 5a along the $Proportion_{API}(\mathcal{O})$ y-axis, depicting the number of unique users that access a given ontology through the BioPortal API.

The scatter plots show that users tend to explore a larger proportion of any given biomedical ontology using the BioPortal WebUI than they query through the BioPortal API. However, certain UMLS terminologies, such as SNOMED CT, NCIt (National Cancer Institute Thesaurus), ICD10 (International Classification of Diseases), MEDDRA (Medical Dictionary for Regulatory Activities) and MEDLINE PLUS stand out as outliers. Almost all classes in these terminologies are accessed using the API, but only a very small number of classes in these terminologies are accessed using the BioPortal WebUI, even though more than $1,000$ unique users have explored these terminologies using the WebUI during the investigated time period. These terminologies are popular, and are used for several biomedical purposes, such as cancer knowledge management, medical records annotation and drug–adverse reaction association discovery [41]. Automated methods, developed by biomedical informatics researchers, query the BioPortal API exhaustively for these purposes. However, the size of these ontologies (10,000–400,000 classes) might make exploration using the WebUI's indented tree visualization difficult.

On the other hand, much smaller subsets of the popular OBO Foundry ontologies, such as the Gene Ontology (GO) and ChEBI, were accessed through the API ($\approx$ 10%). Unlike GO, however, a large proportion of the ChEBI ontology was accessed using the WebUI. Approximately $1,000$ users have accessed these ontologies through BioPortal during the time period between January 2013 and June 2016. Note that several ontology exploration and browsing tools exist for GO and ChEBI, which may serve as alternatives to BioPortal.

Using the yellow–blue color scale of the ontology nodes, we observe that most of these ontologies have a very small Spearman Correlation statistic for *WebUI Access versus API Access* (< 0.5). That is, the set of classes that are accessed using the WebUI may be drastically different from the set of classes that are accessed using the API resource. This finding likely reflects the different goals of users who explore an ontology compared with those who query the ontology using the BioPortal API.

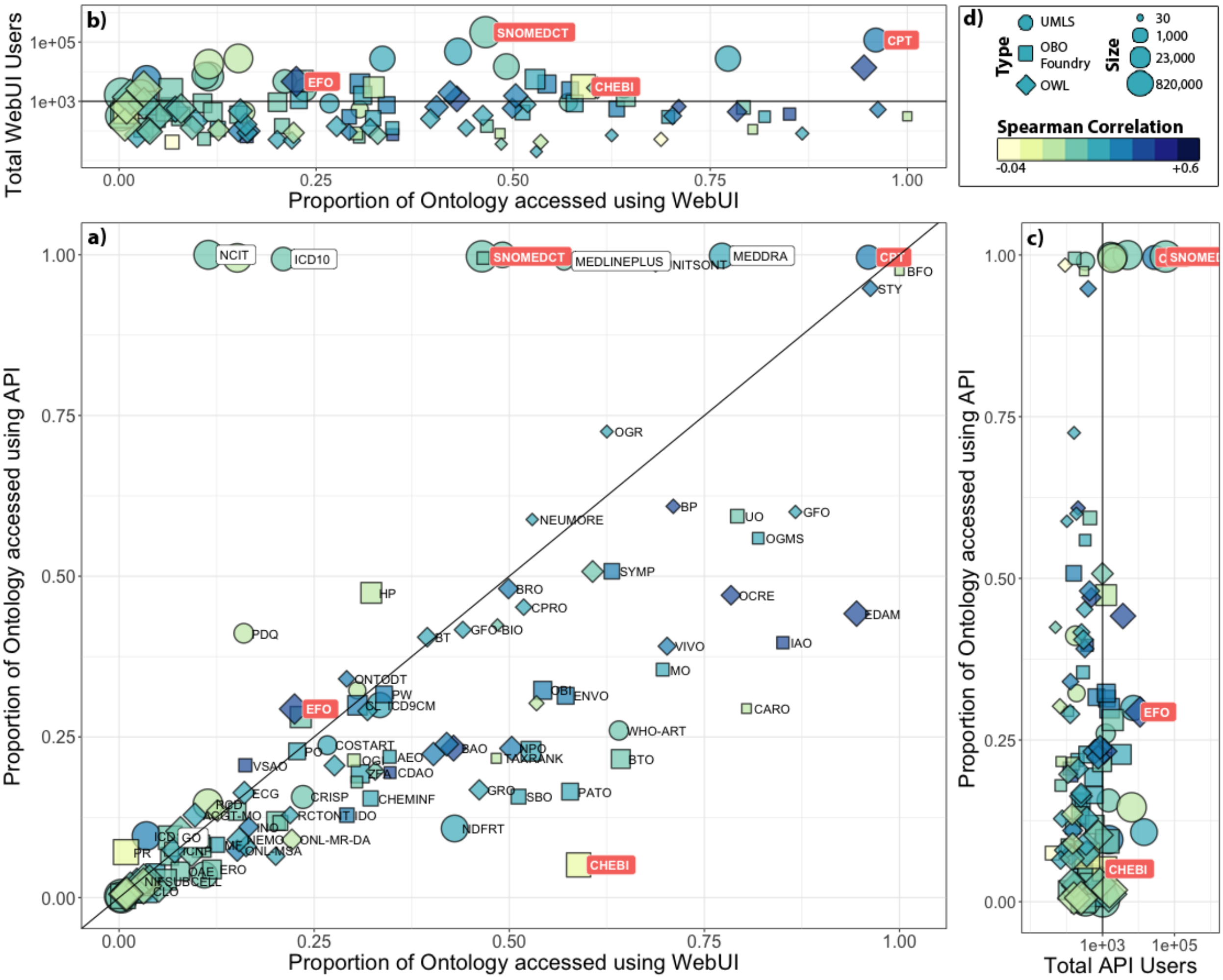


Figure 5: **Visualizing ontology access statistics. a)** The main scatter plot indicates the proportion of the ontology that is accessed using the BioPortal WebUI versus the proportion of ontology that is accessed using the BioPortal API. Each scatter point is an ontology. The size of the point is proportional to the total number of classes in the ontology. The Spearman correlation for the ontology's *WebUI Access versus API Access* comparison is mapped to a yellow–blue color scale (**d**) and used to color the corresponding scatter point. The shape of the ontology depends on whether the ontology is listed under OBO Foundry (square), UMLS (circle) or OWL (diamond). **b)** The horizontal scatter plot indicates the total number of unique IP requests made using the WebUI between 2013–2016. **c)** The vertical scatter plot indicates the total number of unique IP requests made using the API between 2013–2016. Our selected ontologies are highlighted using red-boxed labels.

5.2. Similarity & Overlap of Exploration, Query and Reuse Data across Ontologies

To better understand the findings of Section 5.1, we investigate the distribution of Spearman Correlation and Jaccard Similarity statistics (Figure 6) across the three different comparisons: *i)* Reuse versus API Access, *ii)* Reuse versus WebUI Access, *iii)* WebUI Access versus API Access.

We compute adjusted p-values by differentiating the distributions of these statistics for the different comparisons. To deal with the multiple comparisons problem, we use ANOVA (analysis of variance) summaries and Tukey HSD (Honest Significant Differences) post-hoc tests over those summaries to compute these adjusted p-values. The differences in the distributions, as well as adjusted p-values are presented in Table 2.

The Spearman Correlation statistics (Figure 6a) for the *WebUI Access versus API Access* comparison are significantly higher than they are for the *Reuse versus WebUI Access* and *Reuse versus API Access*. The Spearman correlation statistics for the

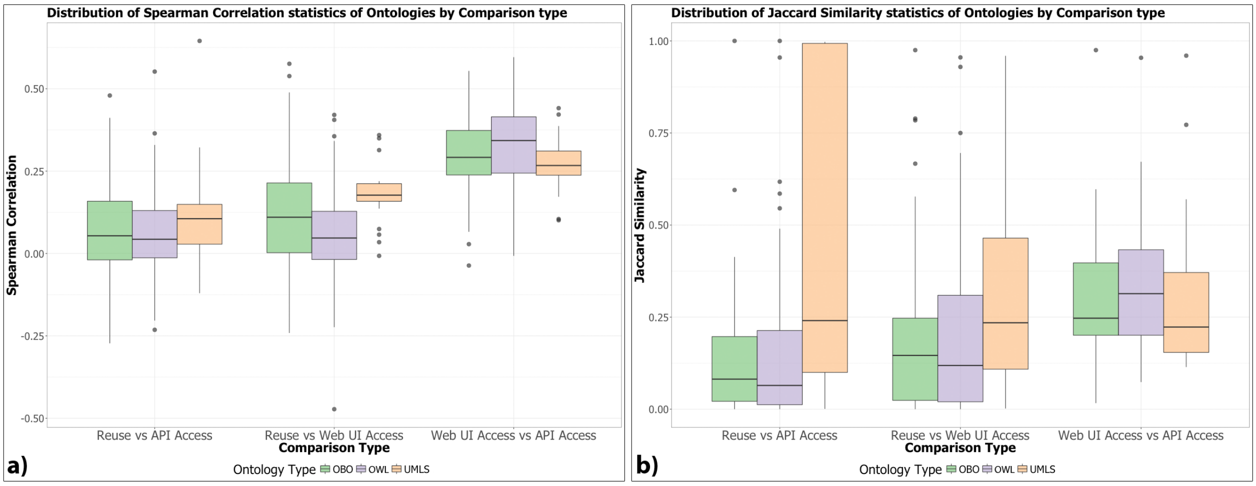


Figure 6: **Distributions of statistics across ontologies.** **a)** Spearman Correlation and **b)** Jaccard Similarity. These distributions are computed for the three main comparisons: **1)** Reuse versus API Access, **2)** Reuse versus WebUI Access, and **3)** WebUI Access versus API Access. These distributions are grouped according to the ontology type—OBO Foundry (green box plot), UMLS (orange box plot) or OWL (purple box plot). The statistics for the 'WebUI Access versus API Access' comparison are generally higher than other comparisons.

Statistic type	**Distribution pair**	**Mean difference**	**Adjusted *p*-value**
Spearman Correlation	Reuse vs WebUI – Reuse vs API	0.01418511	0.4199016
	WebUI vs API – Reuse vs API	0.26334301	0.0000000
	WebUI vs API – Reuse vs WebUI	0.24915790	0.0000000
Jaccard Similarity	Reuse vs WebUI – Reuse vs API	0.01094695	0.8008144
	WebUI vs API – Reuse vs API	0.20231318	0.0000000
	WebUI vs API – Reuse vs WebUI	0.19136624	0.0000000

Table 2: Significance of differences between the distributions of statistics for different comparisons.

first two comparisons (*Reuse versus API Access*, and *Reuse versus WebUI Access*) are generally $<$ 0.1 for most ontologies. This finding indicates that users on BioPortal are less likely to explore or query those classes in a given ontology that are reused by other ontologies. This finding disproves our first research hypothesis which expected that reused classes are also frequently explored and queried. Hence, the number of times a particular class is reused may not be a good indicator of popularity, in terms of exploration and programmatic query access, for most users on BioPortal. Note that the Spearman Correlation statistic for the *Reuse versus WebUI Access* comparison in UMLS terminologies (mean diff. $= 0.11$, $p = 10^{-5}$) and OBO Foundry ontologies (mean diff. $= 0.07$, $p = 2.1 \times 10^{-4}$) is significantly higher than in other OWL ontologies.

Moreover, the Jaccard Similarity statistics (Figure 6b) for the *Reuse versus WebUI Access* and the *Reuse versus API Access* are also very small (≈ 0.25 for UMLS terminologies and ≈ 0.1 for OBO Foundry and other ontologies). Similar to Spearman correlation statistics, the Jaccard Similarity statistics for the *WebUI Access versus API Access* comparison are significantly higher than the statistics for the *Reuse versus API Access* and the *Reuse versus WebUI Access* comparisons. This result further strengthens the rejection of our first research hypothesis. UMLS terminologies exhibit larger correlation and similarity statistics than other ontologies, as we have considered reuse to occur through Concept Unique Identifier (CUI) mappings between similar classes, generated and updated *a posteriori* after the terminologies have been created [8].

However, in general, it should be noted that the Spearman Correlation and the Jaccard Similarity statistics for the *WebUI Access versus API Access* comparison are still very small. This result disproves our second hypothesis, stating that users explore and programmatically query the same classes.

Hence, we may expect that the classes explored using the BioPortal WebUI are less likely to also be queried using the BioPortal API, and vice versa.

5.3. Providing a Visual Perspective on Exploration, Query and Reuse Statistics

To learn more about how users access ontologies using the BioPortal WebUI and API, as well as how they reuse classes from these ontologies, we visualize large amounts of interaction data using PolygOnto. We use as examples four prominent ontologies from the biomedical domain: *i)* Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), *ii)* Chemical Entities of Biological Interest Ontology (ChEBI), *iii)* Current Procedural Terminology (CPT), and *iv)* Experimental Factor Ontology (EFO). These ontologies have between 10,000 and 350,000 classes. More than 1,000 users accessed these ontologies between 2013–2016 using both the BioPortal WebUI and the BioPortal API. The classes of these ontologies are also frequently reused in other ontologies. CPT and EFO have a high Spearman Correlation statistic for the *WebUI versus API* comparison. The statistics for these ontologies are listed in Table 3.

Figure 7 shows three PolygOnto visualizations that we generated for each of the four ontologies. The three visualizations correspond to the WebUI Access, API Access, and Reuse. We also overlay orange circular nodes over each PolygOnto visualization to depict the total number of users who only explore or query a single class in the ontology at a given maximum depth. In each PolygOnto visualization, the underlying red polygon depicts the hierarchical structure of the visualized ontology. The height of the polygon indicates the maximum depth of the ontological hierarchy (number of layers) and the width of each layer indicates the total number of classes at a given depth from the root class in the ontology. In addition, in the WebUI Access and API Access PolygOnto visualizations, each user interaction pattern is displayed as a distinct blue polygon (Section 4.3) based on the hierarchical location of the classes accessed by the user. In the Reuse PolygOnto visualization, each distinct blue polygon represents an ontology that reuses a set of classes from the visualized ontology. The area of each blue polygon is representative of the number of ontology classes accessed by a particular user or reused by another ontology.

SNOMED CT and CHEBI have more than 20 hierarchical layers, and $\approx 300,000$ and $\approx 54,000$ classes, respectively. The number of classes at each layer in the ontological hierarchy may vary drastically, giving a unique shape to each ontology.

5.3.1. Differentiating User Behaviors on the BioPortal WebUI and the BioPortal API

SNOMED CT stands out as an outlier in Figure 5a in that the proportion (%) of classes queried using the API is almost 100. This observation is also clearly visible in the PolygOnto visualization for SNOMED CT API Access Figure 7a, where some users have queried every class in the 2015 version of the SNOMED CT terminology. The ChEBI ontology stands out in Figure 5b as it has a higher proportion of classes that are accessed using the BioPortal WebUI compared to the BioPortal API.

In general, BioPortal users are more likely to explore single classes of interest using the BioPortal WebUI than they are to query a single class using the BioPortal API, which is evident by the larger orange circular nodes in the WebUI PolygOnto visualizations of Figure 7 (highlighted with orange arrows). As the BioPortal API provides programmatic and unrestricted access to an ontology, we assume that users are likely to access larger portions of a given ontology than they are in the WebUI. For example, users who explore or query more than one class in SNOMED CT on BioPortal exhibit a higher tendency to use the BioPortal API ($mean \approx 34$) than the WebUI ($mean \approx 8$). These findings are also statistically significant ($t = 2.31724, p = 0.0205$). However, we do not observe this behavior across all ontologies, as can be seen for ChEBI and CPT ontologies in Figures 7b and 7c respectively. In CPT, users are exploring more classes using the BioPortal WebUI ($mean \approx 6$) than the API ($mean \approx 4$) ($t = -3.827743, p = 0.00013$). From the PolygOnto visualizations for ChEBI, we observe users accessing a larger portion of the ontology using the WebUI ($mean \approx 53$) than using the API ($mean \approx 9$). However, this finding is not statistically significant. Hence, there may only be a few users who explore larger sets of classes using the WebUI. This can be verified using the PolygOnto visualization, where we see four large and distinct user polygons (highlighted with blue arrows).

We generated PolygOnto visualizations for all ontologies in our dataset. We found that users query a larger portion of UMLS terminologies (MESH, LOINC, MEDDRA) using the BioPortal API (average classes queried $\approx 6 - 300$), whereas they explore more classes of smaller ontologies, or those

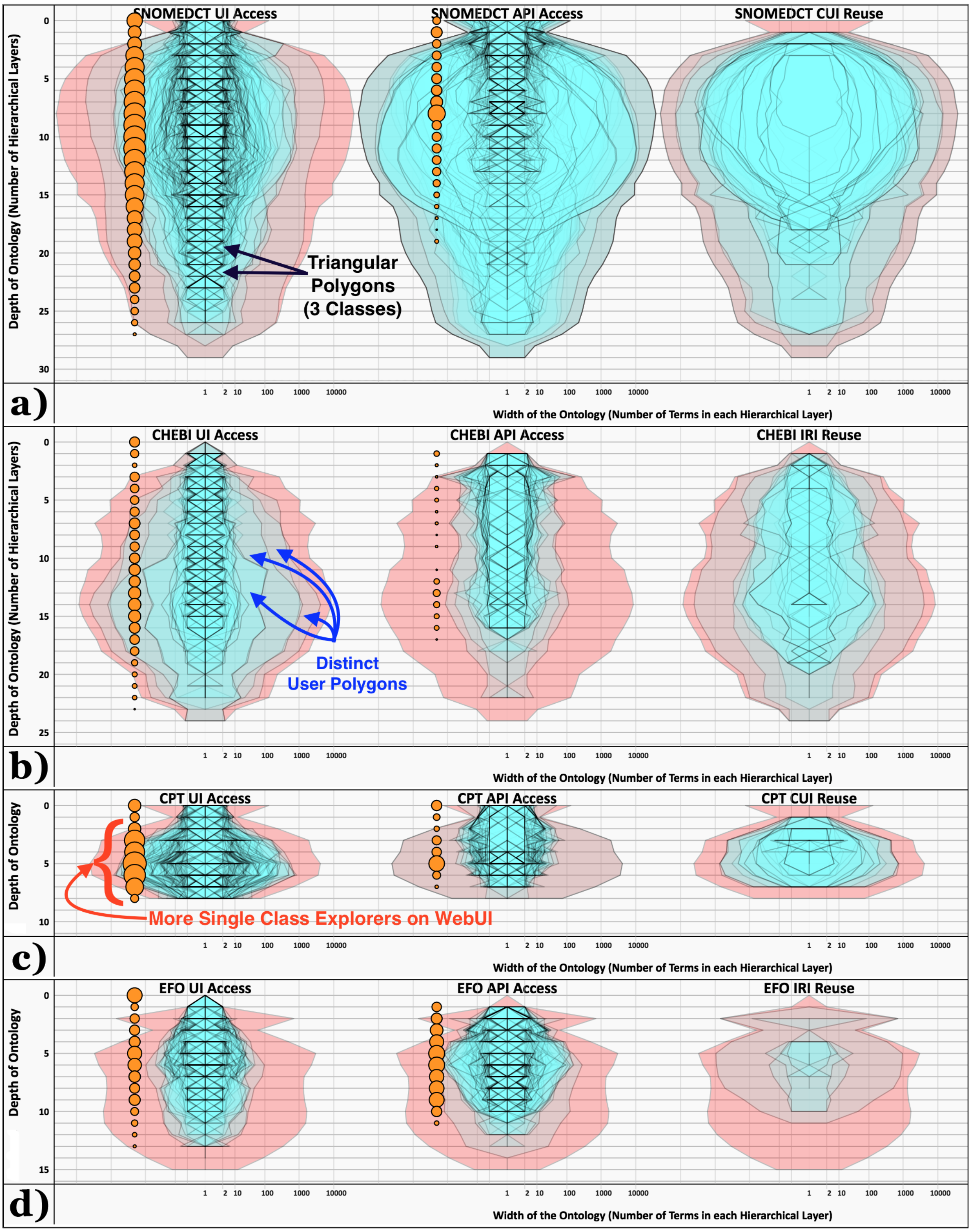


Figure 7: **PolygOnto visualizations.** **a)** Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), **b)** Chemical Entities of Biological Interest (ChEBI), **c)** Current Procedural Terminology (CPT), and **d)** Experimental Factor Ontology (EFO). For each ontology, we show 3 PolygOnto visualizations for WebUI Access, API Access and Reuse.

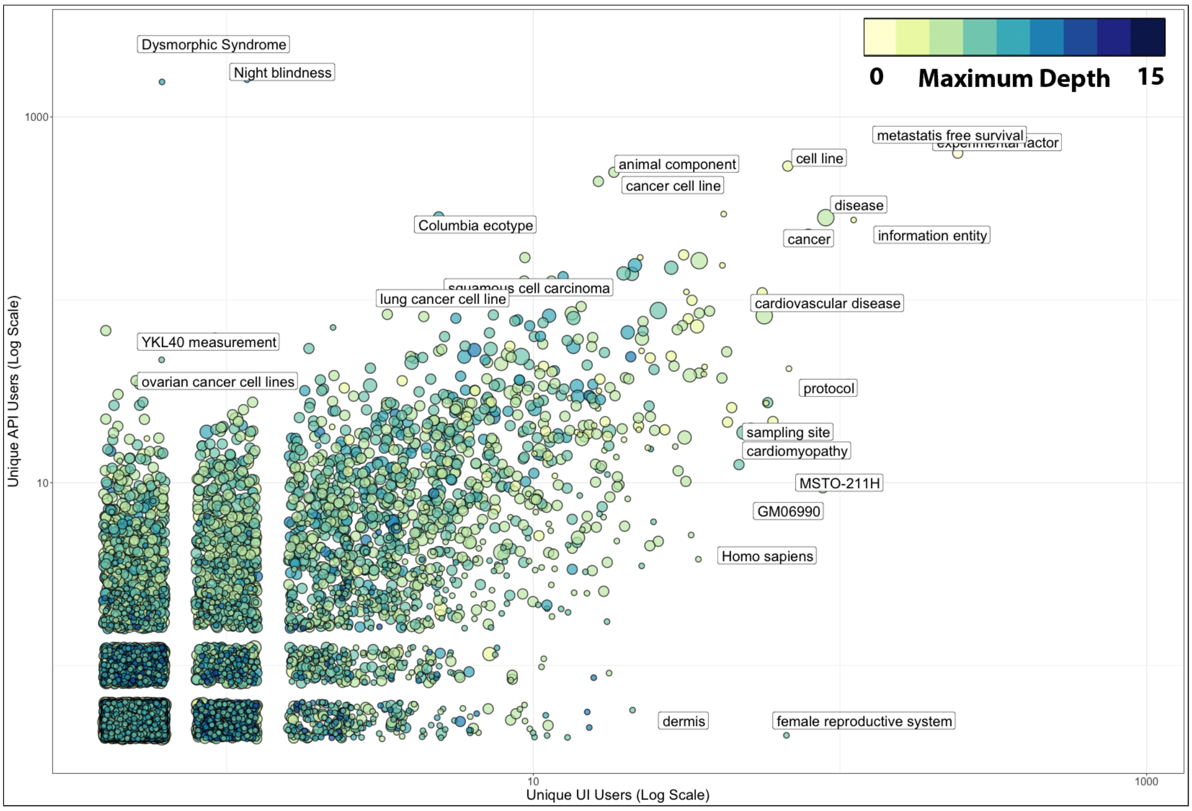


Figure 8: **Scatter plot visualization.** The classes of the Experimental Factor Ontology (EFO) are depicted here as nodes arranged according to the total number of unique users that explore or query the corresponding class using the BioPortal WebUI or the BioPortal API resources. The size of each node is proportional to the number of times the class is reused in other ontologies, and the color of each node is indicative of the **maximum depth** of the class in the ontology (Yellow–Blue color scale). The position of the nodes is jittered to show underlying nodes (e.g., EFO classes unvisited using WebUI or API).

with fewer hierarchical layers ($\approx 5 - 10$ layers) in the WebUI.

5.3.2. Lower Level Classes are Less Explored, Queried & Reused

In Figure 7, we can clearly observe that users rarely explore or query classes in the lower levels of the ontologies (≈ 200–500 classes). We have made the same observation for other ontologies with large hierarchies in BioPortal, such as GO and NCIt. CPT stands out as an exception (Figure 7c and Figure 5b), as the proportion of the explored and queried classes is ≈ 1. It should be noted, that not all leaf classes in an ontology are displayed in the lower levels of a PolygOnto visualization (Figure 4).

This observation can be verified if we generate a scatter plot, such that all classes in the ontology are aligned according to the number of times they have been explored or queried using the BioPortal WebUI and the API respectively (log-scale), and are colored according to their maximum depth attribute. As an example, we generate such a scatter plot for the Experimental Factor Ontology (Figure 8). The positions of the nodes are jittered randomly, so that classes having the same value of unique WebUI and API users are spread across a small region. Four distinct box-shaped regions emerge in the scatter corresponding to the coordinates $(webui = 0, api = 0), (webui = 1, api = 0), (webui = 0, api = 1), (webui = 1, api = 1)$. We can see that the color in these regions tends to be a darker shade of blue, indicating that these regions are composed of classes whose maximum depth attribute is high.

5.3.3. Empirical Usage Patterns Extracted from Exploration, Query & Reuse Data

In Figure 7, we observe triangular polygons along the central axis of the ontology structure for both the WebUI Access and the API Access across all

Ontology $\mathcal{O}$	Size N	Prop$_{\text{UI}}$ %	Prop$_{\text{API}}$ %	Spearman Correlation	Jaccard Similarity	Users WebUI (#)	Users API (#)
SNOMED CT	300,543	46.49	99.76	0.2475	0.4644	215,725	57,109
ChEBI	54,847	58.88	5.08	0.0658	0.0546	3,117	1,135
CPT	13,084	96.00	99.62	0.4412	0.9599	117,130	28,106
EFO	15,294	30.44	29.34	0.5458	0.4312	4,655	10,993

Table 3: Access Statistics for selected ontologies for which the PolygOnto visualizations are presented in this study.

ontologies. These observations indicate exploring and querying patterns that are either *i) Triangles:* 1 parent → 2 child classes, or *ii) Inverted Triangles:* 1 child → 2 parent classes.

We found that the median number of classes a user explores using the WebUI is 3. This fact is easily observed in the PolygOnto visualization, i.e. the smallest triangle or inverted triangle in the PolygOnto visualization indicates 3 classes. More triangular polygons are observed for the WebUI than API Access PolygOnto plots, especially in the lower layers of the class hierarchies. This finding indicates a direct effect of the indented tree visualization that is used in the BioPortal WebUI, which facilitates the exploration of siblings as well as children or parents in the hierarchy. Using the BioPortal API, most users query the lower levels of the hierarchy only when already requesting larger parts of the ontology. This contrasts with our initial expectation that the BioPortal API would be used mainly in downstream applications to query specific classes in the lower levels of the hierarchy (i.e., less abstract classes).

5.4. Comparing Usage in Real-World Applications

To investigate if and to what extent browsing, query and reuse data from BioPortal informs real-world usage of classes, we analyze the use of EFO to annotate GWAS datasets (Section 5.4.1) and the use of the ChEBI ontology for data integration in the Life Sciences Linked Open Data Cloud and annotation of PubChem compounds (Section 5.4.2).

5.4.1. Usage of the Experimental Factor Ontology

We use the methods described in Section 4.1 to compare and correlate user browsing and querying behavior, as well as reuse, for the Experimental Factor Ontology (EFO) in BioPortal, with the usage of EFO in the NHGRI GWAS Catalog. We also apply the PolygOnto visualization method (Section 4.3) over EFO and the NHGRI GWAS Catalog to determine the more useful portion of EFO for annotating GWAS studies, and also to determine the depth of classes that are more commonly used for annotation. Each study in the NHGRI GWAS Catalog is similar to a unique user in these experiments.

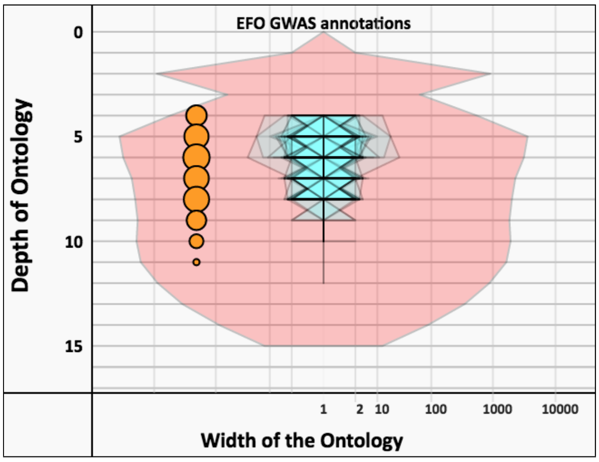


Figure 9: **Visualizing the Experimental Factor Ontology (EFO) and the EFO-annotated traits analyzed in the genome-wide association studies (GWAS) from the NHGRI GWAS Catalog.** Each polygon here is an unique study. The circles represent single class annotations.

Table 4 summarizes the Spearman correlation and the Jaccard similarity statistics of the WebUI Access, API Access and the Reuse data, with respect to the EFO usage data. The Spearman correlation values for all three comparisons (including *Reuse versus Usage*) are in the range of 0.2 – 0.4 (similar to *WebUI Access versus API Access* Spearman correlation statistics presented in Section 5.2). It should be noted that these statistics are observed even though only ≈ 5% of EFO is used to annotate the NHGRI GWAS Catalog studies. Due to the low usage of EFO, the Jaccard similarity scores are lower for these comparisons. After limiting our analysis to only those EFO classes that are used for annotations, we get adjusted Spearman correlation statistics for *WebUI Access versus Usage* and *API Access versus Usage* comparisons that are higher

Ontology $\mathcal{O}$	Dataset/ Scenario	Prop$_{\text{Used}}$ %	Compared Attribute	Jaccard Similarity	Spearman Correlation	Adjusted Sp. Correlation
EFO	GWAS Catalog	5.49	WebUI Access	0.1258	0.2076	0.2963
			API Access	0.1715	0.3997	0.4945
			Reuse	0.1668	0.3240	0.1529
ChEBI	LSLOD Cloud	84.06	WebUI Access	0.5408	0.0029	0.0354
			API Access	0.0369	-0.0696	0.0326
			Reuse	0.7461	-0.1159	0.1017
	PubChem Assays	80.80	WebUI Access	0.5354	0.0407	0.0025
			API Access	0.0310	-0.0978	0.0465
			Reuse	0.3951	-0.0751	0.2148

Table 4: Statistics for selected ontologies and biomedical scenarios that use ontology-based annotations.

still — at ≈ 0.3 and ≈ 0.5 respectively.

In the PolygOnto visualization shown in Figure 9, studies that are annotated with 2 or more EFO classes are represented using polygons, whereas those studies that are annotated with a single EFO class are summarized as the adjacent (orange) circles. We observe that no study has been annotated with classes in the lower levels (≈ 1,000 classes from levels 12–16), or upper levels (that may originate from the Basic Formal Ontology) of the EFO hierarchy. This observation also reflects the results in Section 5.3, where we found that the lower levels of the ontological hierarchy in some of the highly-accessed ontologies are rarely (if at all) explored by users.

Based on the Spearman correlation statistics and PolygOnto visualization, we assert that the EFO classes used for annotations in the NHGRI GWAS Catalog are also somewhat more frequently explored and queried by users through the BioPortal WebUI and BioPortal API respectively, with very similar grouped interaction patterns.

5.4.2. Usage of the Chemical Entities of Biological Interest (ChEBI) Ontology

We analyzed the use of ChEBI ontology in two real-world biomedical application scenarios.

Scenario 1. LSLOD cloud. The first scenario involved the use of ChEBI to annotate or cross-reference schema elements, as well as individual instances, in 12 different RDF data sources in the Life Sciences Linked Open Data (LSLOD) cloud. These sources may use ChEBI-annotated schema elements to structure their data or to cross-reference relevant instances (e.g. compounds, small molecules, chemicals, or drugs) with corresponding classes in the ChEBI ontology. Each polygon in the PolygOnto visualization (Figure 10) refers to a distinct LSLOD source. The PubChem data source [38] is highlighted in green since we further analyze the use of ChEBI in that source in our second scenario.

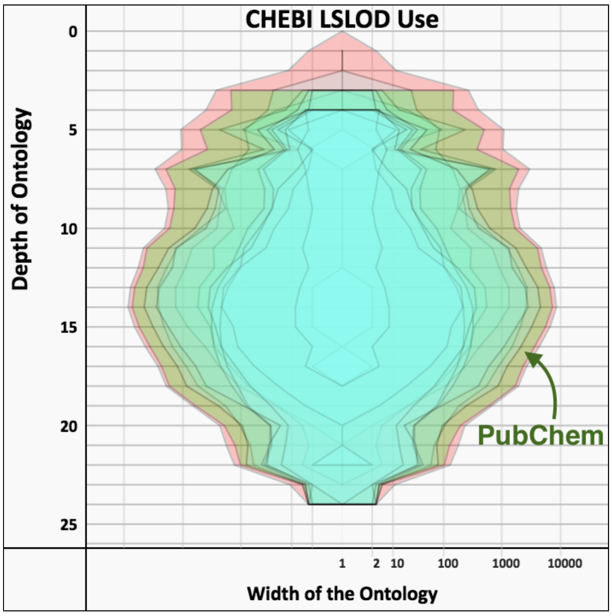


Figure 10: **Visualizing the Chemical Entities of Biological Interest Ontology (ChEBI) and its use for integration in the Life Sciences Linked Open Data (LSLOD) sources** Each polygon here is a distinct LSLOD data source, that uses ChEBI classes at schema and instance levels. The green-colored polygon represents the set of ChEBI classes that are used in the PubChem data source.

Usage of the ChEBI ontology (54,847 classes in the January 2015 version) is quite varied across the sources with the number of classes used for annotations ranging from 16 in BioSamples to 44,305 in PubChem (*median* ≈ 3,125, *std. dev.* ≈ 13,675). This is likely due to the different purposes for which

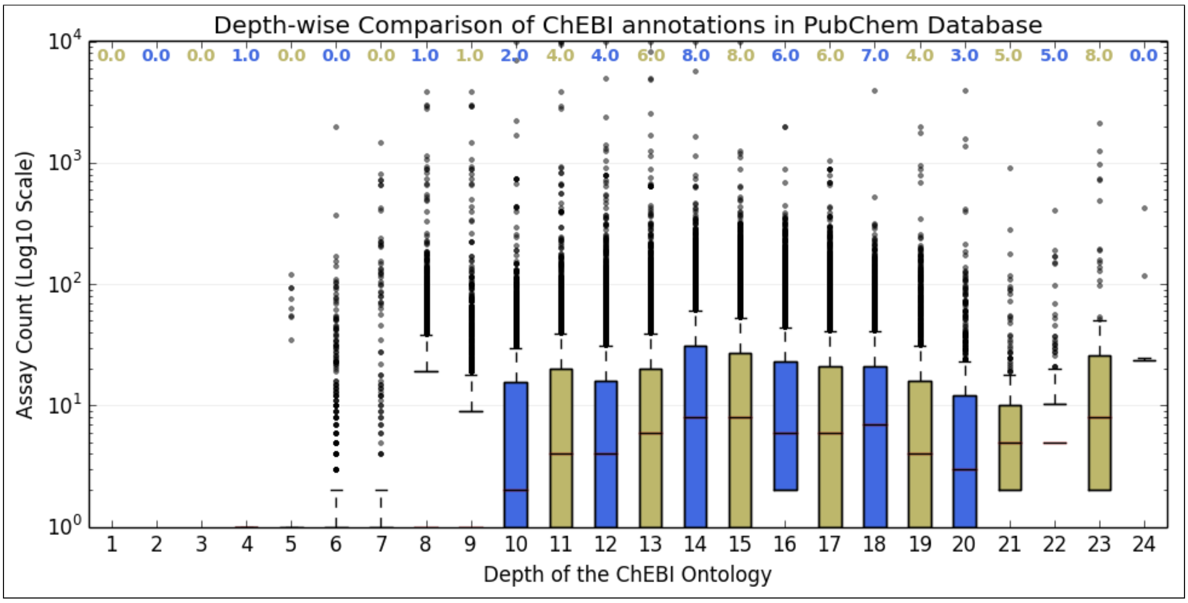


Figure 11: **Visualizing the distributions of PubChem assay counts that use compounds annotated by ChEBI classes.** These distributions are grouped according to the depth of the ChEBI ontology, with each box plot representing the distribution of PubChem assay counts for the set of ChEBI classes at a given ontological depth. The median for each distribution is shown above the plot, and the colors are alternated for visual discernibility. It can be clearly observed that the PubChem database uses ChEBI classes for annotation from the lower levels of the ontological hierarchy.

the LSLOD sources have been developed. For example, DrugBank is a database of drugs and its molecular properties and may only have cross references to relevant drug-based classes present in the ChEBI ontology [39]. The varying shapes of the different polygons in the PolygOnto visualization depict this observation.

Compared with Figure 7b, which visualized WebUI Access, API Access, and Reuse, for the ChEBI ontology, Figure 10 indicates that classes in the lower layers of the ChEBI hierarchy are actually used for annotation in several sources in the LSLOD cloud. However, the classes in the upper layers of the ChEBI hierarchy ($\approx 1,000$ classes) are rarely used in these sources, and may be deemed too abstract for actual biomedical research. Most LSLOD sources in our study start using classes from Layer 3 of the ontological hierarchy. It should be noted that, in ChEBI, the upper layer classes are not reused from other upper level ontologies, such as the Basic Formal Ontology.

Scenario 2. PubChem compound annotations. We observe a similar usage behavior in a second scenario when we analyzed the usage of ChEBI classes to annotate the compounds used in PubChem biological assays. In Figure 11, we visualize the distributions of assay count data (i.e. number of biological assays that use a compound annotated with a particular ChEBI class) and group the distributions according to the ontological depth of ChEBI. We show the median number of assays that use a particular ChEBI-annotated compound on the top of each distribution. Further, we see that classes, located after Layer 10 of the ChEBI ontological hierarchy, are significantly used for the annotation of PubChem compounds. The median number of biological assays that use a particular compound annotated with a ChEBI class from Layers 11–23, ranges from 4–8 assays, compared to 0–2 for ChEBI classes in Layers 1–10.

Unlike the results for EFO (Section 5.4.1) scenario, we observe a minimal or negative correlation between the use of ChEBI classes in the LSLOD cloud and the browsing and querying behaviors of the BioPortal users, as well as reuse data for the ChEBI ontology (Table 4). Similarly, the correlation statistics obtained for the second ChEBI scenario, between the use of ChEBI to annotate compounds in the PubChem data source and the ChEBI browsing, querying and reuse data, are also minimal and negative. The adjusted Spearman correlation statistics slightly increase, especially when comparing the usage attributes to the reuse attribute.

The contrasts between usage and the other attributes might be the reason for the negative Spearman correlation statistics (i.e. no upper layer classes used in both scenarios). These statistics increase on adjusting our ChEBI class sets to include only those classes that are used in the scenarios (Table 4). However, because these correlation statistics are much lower compared to our EFO usage scenario, our third hypothesis is refuted for the two ChEBI-based usage scenarios.

6. Discussion

We have investigated how researchers explore, query, use and reuse ontologies in BioPortal by conducting an exploratory visual empirical analysis of large-scale interaction data. We used scatter plots (Figure 5 and Figure 8) to learn whether users more frequently explore ontologies via the BioPortal WebUI or query them over the API. We used box plots to analyze Spearman correlation coefficients and Jaccard similarity (Figure 6) for investigating correlation and overlap among information consumption strategies across different interfaces of BioPortal and ontologies. We visually inspected exploration, query and reuse strategies of thousands of different users for four different ontologies (Figure 7) using PolygOnto visualizations. Finally, we analyzed the usage of EFO for annotating the GWAS Catalog, and the usage of ChEBI for data integration through the LSLOD cloud and the annotation of compounds used in biological assays stored in PubChem database (Section 5.4).

6.1. Exploring, Querying and Reuse

We found that the Spearman Correlation statistics for all ontologies for the *Reuse versus WebUI Access* and *Reuse versus API Access* comparisons are significantly lower than the Spearman correlation statistics for the *WebUI Access versus API Access* comparison. Both the Spearman correlation and the Jaccard Similarity statistics are generally lower with a median of ≈ 0.1. Hence, from a BioPortal perspective, we can assert that the classes that are reused the most in other ontologies are not the same classes that BioPortal users explore or query more often. This insight also suggests a negative answer to our first research question: **RQ1: Do BioPortal WebUI exploration and API querying inform reuse?**

A possible explanation for this finding is that most reused classes are at a more abstract level, or originate from upper level ontologies. Such classes may not be useful to biomedical researchers in their applications (hence less BioPortal exploration and API querying). The main purpose of upper-level ontology class reuse is to extend the interoperability among ontologies. Researchers can directly search for their classes of interest by exploring an ontology using the BioPortal WebUI or querying the API. Note that both the Basic Formal Ontology (BFO) and the Semantic Types Ontology (STY), which are used as upper-level ontologies in OBO and UMLS ontologies, have higher proportions of classes explored and queried, and exhibit higher values for the Spearman and Jaccard statistics (Figure 5b).

OBO and OWL ontologies have lower Spearman correlation and Jaccard similarity statistics for the *Reuse versus WebUI Access* comparison when compared to the UMLS terminologies. The ontologies in the former group use the IRI method to reuse classes, whereas for the latter group, domain experts map similar classes *a posteriori* using the same CUI. The fact that the CUI mappings are generated *a posteriori* may help identify more cases of reuse than does IRI reuse mapping (which typically happens at the time the ontology is authored). OBO ontologies have significantly higher Spearman correlation statistics than OWL ontologies, which might indicate the importance of guidelines for reuse.

Knowing that the upper layers of the ontologies are of less interest to biomedical researchers could prove useful to BioPortal developers. For example, a filtering option could hide all the abstract classes of BFO or other upper level ontologies when they appear as part of an ontology.

We need to carry out further investigations to understand why certain classes were queried using the API but were never explored using the WebUI during the same period of time— and vice versa. For example, the classes `Night blindness` and `Dysmorphic syndrome` (Figure 8) are queried by more than 1,000 unique users via the BioPortal API, but are never explored through the WebUI. Such specific class queries are observed for all ontologies, and we may require domain expertise to decipher the importance of these. Even though *Web UI Access versus API Access* comparisons exhibit slightly higher correlation values, the classes that are explored and queried exhibit minimal similarity, suggesting a negative answer to our second research question: **RQ2: Do BioPortal WebUI exploration and API querying correlate?**

6.2. Ontology Use in Real-World Applications

When inspecting Figure 7, we see that classes in the lower layers of the class hierarchies are rarely explored via the BioPortal WebUI and rarely queried using the BioPortal API. This observation may lead to further questions regarding the utility of more specific classes in ontologies, if they are not of interest to domain users (Figure 7, Figure 9).

However, in Section 5.4.2, we observed that specialized classes in the lower layers of the ChEBI ontological hierarchy are used significantly more in two real-world usage scenarios (Section 5.4.2). These observed contrasts between user interactions with different consumption strategies (WebUI and API), and actual usage may open new research avenues to determine whether the conventional interaction methods (e.g. Indented Tree visualization) are actually fulfilling user needs, and whether new methods need to be developed to replace them.

In the case of the Experimental Factor Ontology, actual usage yielded higher Spearman Correlation statistics and similar usage patterns, when compared to users' browsing and querying behaviors. Classes in lower levels of the EFO hierarchy are not used in the annotation of GWAS studies, since GWAS studies are generally conducted using a large cohort of patients for achieving adequate statistical power [43]. Hence, if a more specific trait (i.e., one that appears lower in the EFO hierarchy) is examined, the investigators of the study may have fewer patients to conduct a GWAS. The high correlation statistics observed between the usage for GWAS annotations and other attributes (access or reuse) does not indicate causality. It is impossible to determine if the classes are increasingly explored, queried or reused because they were used for GWAS annotations, or whether GWAS studies were conducted because the traits were popularly browsed or searched.

We cannot definitely answer our third research question (**RQ3: Do BioPortal WebUI exploration and API querying inform usage?**) since different usage scenarios may exhibit different usage patterns. To further strengthen the generalization of our results, more analyses need to be carried out on datasets generated by other applications. To encourage other researchers to continue this research, we have processed and published the user interaction datasets used in this study at: `http://onto-apps.stanford.edu/bionic` under the Creative Commons CC-BY-NC-SA license [11]. We have also released the source code used to generate the statistics and the visualizations in this paper at: `https://github.com/maulikkamdar/PolygOnto`. We will soon provide a packaged solution for generating PolygOnto visualizations given an ontology and relevant datasets.

6.3. Limitations and Future Work

Note that this analysis is limited to BioPortal. Some users might prefer to explore or query certain ontologies using other platforms (e.g., the AmiGO browser [44] for Gene Ontology, or Ontology Lookup Service [45] for ChEBI), which provide custom tailored or more specialized interfaces for the task of browsing and querying ontologies. Hence, we may register fewer requests on BioPortal for these ontologies. Moreover, the user interactions with BioPortal WebUI may be biased by the current Indented Tree visualization used for exploration.

There are certain limitations to our empirical analysis. By generating a high-level abstract representation using PolygOnto visualizations (Figure 7), we lose information, such as the class-level counts (i.e., number of times a particular class was accessed in an ontology), that can be easily visualized in a scatter plot (Figure 8). We also lose information regarding the location of each class in the class hierarchy by generalizing it to the maximum depth attribute. In certain cases, the leaf classes in a given ontology might be located in the middle layers of the ontology because there may be classes with a higher maximum depth attribute.

For future work, we plan to evaluate the utility of PolygOnto for ontology engineers, ontology repository developers, and biomedical researchers by conducting a large-scale user study. We have created an interactive Web-based visualization[2] for such an evaluation, where users may use the scatter plot visualizations to explore count-based data for specific ontology classes, and may use the PolygOnto visualization to visually investigate higher-level abstract patterns as demonstrated in this work.

7. Conclusion

We have conducted an empirical analysis of request, query and reuse data of large-scale ontologies on BioPortal and demonstrated how several

[2] `http://onto-apps.stanford.edu/vision/`

different visualization techniques can be used to explore, analyze and improve our understanding of how users interact with biomedical ontologies. While ontology browsing, ontology querying and ontology reuse seem to minimally align with each other in our datasets, similarities and differences in the patterns of user exploration, querying and reuse can be observed through our visualization methods. However, these patterns may diverge significantly when they are compared to real-world research use cases in the biomedical domain. To the best of our knowledge, this is one of the largest studies to investigate how users interact with large biomedical ontologies through different modes and for different purposes and downstream applications. Our analyses and methods open new avenues for research into user interactions with ontologies, and will serve as a foundation for future research into the development of intelligent interfaces for ontology exploration and querying.

Acknowledgments

The authors acknowledge Manuel Salvadores and Alex Screnchuk for providing a triplestore dump of BioPortal ontologies. The authors also acknowledge other members of the Protégé Group and the National Center for Biomedical Ontology for their input and evaluation. The authors would like to thank Katharine Shafer Miller for copy-editing the manuscript and providing useful feedback. This work is supported in part by grants GM086587, GM103316, and U54-HG004028 from the US National Institutes of Health.

References

[1] O. Bodenreider, Biomedical ontologies in action: role in knowledge management, data integration and decision support, Yearbook of medical informatics (2008) 67.

[2] D. L. Rubin, et al., Biomedical ontologies: a functional perspective, Briefings in bioinformatics 9 (1) (2008) 75–90, dOI:10.1093/bib/bbm059.

[3] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, Nucleic acids research 36 (suppl 1) (2008) D344–D350.

[4] P. L. Whetzel, 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 39 (suppl 2) (2011) W541–W545, dOI:10.1093/nar/gkr469.

[5] S. Walk, L. Esín-Noboa, D. Helic, M. Strohmaier, M. A. Musen, How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs (2017) 775–784.

[6] B. Fu, N. F. Noy, M.-A. Storey, Indented tree or graph? a usability study of ontology visualization techniques in the context of class mapping evaluation, in: International Semantic Web Conference, Springer, 2013, pp. 117–134.

[7] J. Nair, et al., The BioPortal Import Plugin for Protégé, in: Proceedings of the 2nd International Conference on Biomedical Ontology, Vol. Vol-833, CEUR-WS, 2011.

[8] M. R. Kamdar, T. Tudorache, M. A. Musen, A systematic analysis of term reuse and term overlap across biomedical ontologies, Semantic Web (Preprint) (2016) 1–19.

[9] M. R. Kamdar, T. Tudorache, M. A. Musen, Investigating term reuse and overlap in biomedical ontologies., Proceedings of the International Conference on Biomedical Ontology 2015 (2015) 42–46.

[10] N. Sioutos, et al., NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information, Journal of biomedical informatics 40 (1) (2007) 30–43, dOI:10.1016/j.jbi.2006.02.013.

[11] M. R. Kamdar, S. Walk, T. Tudorache, M. A. Musen, Bionic: A catalog of user interactions with biomedical ontologies, in: The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, 2017.

[12] M. Strohmaier, S. Walk, J. Pöschko, D. Lamprecht, T. Tudorache, C. Nyulas, M. A. Musen, N. F. Noy, How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects, Web Semantics: Science, Services and Agents on the World Wide Web 20 (0).

[13] S. M. Falconer, T. Tudorache, N. F. Noy, An analysis of collaborative patterns in large-scale ontology development projects., in: M. A. Musen, O. Corcho (Eds.), K-CAP, ACM, 2011, pp. 25–32.

[14] C. Debruyne, Q. Reul, R. Meersman, Gospl: Grounding ontologies with social processes and natural language, in: 2010 Seventh International Conference on Information Technology, IEEE, 2010, pp. 1255–1256.

[15] S. Van Laere, R. Buyl, M. Nyssen, A Method for Detecting Behavior-Based User Profiles in Collaborative Ontology Engineering, in: On the Move to Meaningful Internet Systems: OTM 2014 Conferences, Springer, 2014, pp. 657–673.

[16] S. Van Laere, R. Buyl, M. Nyssen, C. Debruyne, Detecting user profiles in collaborative ontology engineering using a user's interactions, Journal on Data Semantics (2017) 1–12.

[17] M. Vigo, C. Jay, R. Stevens, Constructing conceptual knowledge artefacts: activity patterns in the ontology authoring process, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3385–3394.

[18] H. Wang, T. Tudorache, D. Dou, N. F. Noy, M. A. Musen, Analysis of user editing patterns in ontology development projects, in: On the Move to Meaningful Internet Systems: OTM 2013 Conferences, Springer, 2013, pp. 470–487.

[19] S. Walk, P. Singer, M. Strohmaier, D. Helic, N. F. Noy, M. A. Musen, How to apply markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects, International Journal of Human-Computer Studies 84 (2015) 51 – 66.

[20] S. Walk, P. Singer, M. Strohmaier, T. Tudorache, M. A. Musen, N. F. Noy, Discovering beaten paths in col-

laborative ontology-engineering projects using markov chains, Journal of Biomedical Informatics 51 (2014) 254–271.

[21] S. Walk, P. Singer, L. E. Noboa, T. Tudorache, M. A. Musen, M. Strohmaier, Understanding how users edit ontologies: Comparing hypotheses about four real-world projects, in: The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I, 2015, pp. 551–568.

[22] C. Pesquita, F. M. Couto, Predicting the extension of biomedical ontologies, PLoS Comput Biol 8 (9) (2012) e1002630.

[23] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, E. Giannopoulou, Ontology visualization methods—a survey, ACM Computing Surveys (CSUR) 39 (4) (2007) 10.

[24] N. F. Noy, et al., Creating semantic web contents with Protégé-2000, IEEE intelligent systems 16 (2) (2001) 60–71.

[25] S. Lohmann, S. Negru, F. Haag, T. Ertl, Vowl 2: user-oriented visualization of ontologies, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2014, pp. 266–281.

[26] M. R. Kamdar, et al., ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research, Journal of biomedical informatics 47 (2014) 112–130, dOI:10.1016/j.jbi.2013.10.001.

[27] E. H. Baehrecke, N. Dang, K. Babaria, B. Shneiderman, Visualization and analysis of microarray and gene ontology data with treemaps, BMC bioinformatics 5 (1) (2004) 84.

[28] A. Bosca, D. Bonino, P. Pellegrino, Ontosphere: more than a 3d ontology visualization tool., in: Swap, Citeseer, 2005.

[29] H.-J. Schulz, S. Hadlak, H. Schumann, The design space of implicit hierarchy visualization: A survey, IEEE transactions on visualization and computer graphics 17 (4) (2011) 393–411.

[30] B. Johnson, B. Shneiderman, Tree-maps: A space-filling approach to the visualization of hierarchical information structures, in: Proceedings of the 2nd conference on Visualization'91, IEEE Computer Society Press, 1991, pp. 284–291.

[31] J. L. Hintze, R. D. Nelson, Violin plots: a box plot-density trace synergism, The American Statistician 52 (2) (1998) 181–184.

[32] P. Kampstra, et al., Beanplot: A boxplot alternative for visual comparison of distributions, Journal of statistical software 28 (1) (2008) 1–9.

[33] S. Haroz, R. Kosara, S. L. Franconeri, Isotype visualization: Working memory, performance, and engagement with pictographs, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM, 2015, pp. 1191–1200.

[34] D. Winter, B. Vinegar, H. Nahal, R. Ammar, G. V. Wilson, N. J. Provart, An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets, PloS one 2 (8) (2007) e718.

[35] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., The nhgri gwas catalog, a curated resource of snp-trait associations, Nucleic acids research 42 (D1) (2014) D1001–D1006.

[36] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270, dOI:10.1093/nar/gkh061.

[37] B. Smith, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nature biotechnology 25 (11) (2007) 1251–1255, dOI:10.1038/nbt1346.

[38] E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, Pubchem: integrated platform of small molecules and biological activities, Annual reports in computational chemistry 4 (2008) 217–241.

[39] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, Drugbank: a comprehensive resource for in silico drug discovery and exploration, Nucleic acids research 34 (suppl_1) (2006) D668–D672.

[40] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, T. E. Klein, Pharmgkb: the pharmacogenetics knowledge base, Nucleic acids research 30 (1) (2002) 163–165.

[41] M. R. Kamdar, M. A. Musen, Phlegra: Graph analytics in pharmacology over the web of life sciences linked open data, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2017, pp. 321–329.

[42] A. Gelman, J. Hill, Data analysis using regression and multilevel/hierarchical models, Cambridge university press, 2006.

[43] E. P. Hong, J. W. Park, Sample size and statistical power calculation in genetic association studies, Genomics & informatics 10 (2) (2012) 117–122.

[44] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, W. P. W. Group, et al., Amigo: online access to ontology and annotation data, Bioinformatics 25 (2) (2009) 288–289.

[45] R. G. Côté, P. Jones, R. Apweiler, H. Hermjakob, The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries, BMC bioinformatics 7 (1) (2006) 97.