BiOnIC: A Catalog of User Interactions with Biomedical Ontologies

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

Stanford Center for Biomedical Informatics Research, Stanford University
{maulikrk, walk, tudorache, musen}@stanford.edu

Abstract. BiOnIC is a catalog of aggregated statistics of user clicks, queries, and reuse counts for access to over 200 biomedical ontologies. BiOnIC also provides anonymized sequences of classes accessed by users over a period of four years. To generate the statistics, we processed the access logs of BioPortal, a large open biomedical ontology repository. We publish the BiOnIC data using DCAT and SKOS metadata standards. The BiOnIC catalog has a wide range of applicability, which we demonstrate through its use in three different types of applications. To our knowledge, this type of interaction data stemming from a real-world, large-scale application has not been published before. We expect that the catalog will become an important resource for researchers and developers in the Semantic Web community by providing novel insights into how ontologies are explored, queried and reused. The BiOnIC catalog may ultimately assist in the more informed development of intelligent user interfaces for semantic resources through interface customization, prediction of user browsing and querying behavior, and ontology summarization. The BiOnIC catalog is available at: http://onto-apps.stanford.edu/bionic.

Keywords: ontology exploration, reuse, user behavior, log analysis

1 Understanding User Behavior by Analyzing Access Logs

Over the past decade, ontologies have proliferated in the biomedical domain. BioPortal[1]—an open online repository of biomedical ontologies [7]—hosts over 550 ontologies to date. Biomedical researchers use these ontologies to drive a wide-range of biomedical applications [2]. In the first half of 2016 alone, more than 215,000 unique IP addresses submitted 2.52 million requests to access ontologies hosted on BioPortal. Biomedical ontologies often contain thousands of entities, are highly specialized, and they are very expensive to develop. To better serve the ontology development and consumer communities, it is crucial for Semantic Web researchers to gain insights into how these ontologies are explored, queried and reused.

Users may access the content of BioPortal ontologies in two ways: *i)* the BioPortal Web interface (further referred to as *WebUI*) to explore the ontologies, and *ii)* the REST API[2] to query the ontology content programmatically. All access to BioPortal, either via

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

[2] http://data.bioontology.org/documentation

the WebUI or through the API, is captured in the BioPortal Apache access logs, which record the request URL, the time of request, and the IP address of the requestor. We have previously used the information stored in these access logs to identify categories of users accessing ontologies in BioPortal [6], to study ontology reuse [4], and to visualize ontology exploration and query patterns [5].

In this resource paper, we present BiOnIC–the Catalog of User Interactions with Biomedical Ontologies. The catalog contains two types of datasets—*i)* aggregated statistics of user clicks, queries and reuse counts for all classes in BioPortal ontologies, and *ii)* anonymized sequences of user interactions with BioPortal ontologies. The catalog currently contains the anonymized data for access to 255 BioPortal ontologies collected between 2013–2016. The BiOnIC catalog is freely available under the Creative Commons (CC) BY-NC-SA license at: `http://onto-apps.stanford.edu/bionic`.

In the following sections, we describe the vocabulary schema (Section 3), the VisIOn Web application for visualizing the catalog datasets (Section 4), and three applications that rely on the BiOnIC catalog, as well as further application types that would benefit from the BiOnIC catalog (Section 6).

2 BiOnIC Datasets Generation

Filtering the BioPortal logs. We collected all WebUI and API calls from the BioPortal Apache access logs submitted between January 2013–June 2016. To obtain a list of actual user actions, we filtered these logs (e.g., removing robot calls, removing invalid calls) using the filtering methods we developed previously [6]. After the filtering, we obtained 5.4 million WebUI clicks and 67.2 million API queries. For every single request we extracted a valid BioPortal ontology identifier and a class IRI.

Filtering the ontologies. We identified a set of 255 BioPortal ontologies from these logs whose classes were reused in other ontologies. To extract statistics about classes in each of these ontologies (e.g., reuse and sibling count), we pre-processed the ontologies using the methods described in Kamdar et al. [4].

Computing class counts. For each class in each ontology, we aggregated the total and the unique number of users that clicked or queried the class using the WebUI or the API, respectively. We also computed the number of ontologies that reuse each class, and other structural characteristics, such as, the number of sibling classes, direct superclasses, direct sub-classes, and maximum depth of the class in the class hierarchy.

Computing the sequence of user actions. For each ontology, we generated ordered sequences of classes accessed by each user, as well as their associated relative timestamps. While it is possible that a group of users may have used the same IP address to access a set of classes in an ontology, or one user may explore or query classes present in two or more BioPortal ontologies simultaneously, publishing a unique sequence dataset for each ontology facilitates a simple organizational structure for these sequences.

Anonymizing the user data. We converted all logged IP addresses to unique user identifiers, consisting of a randomized string and an integer, encoded with the SHA-224 hashing algorithm [3]. We also converted all absolute timestamps to a value relative to the start timestamp, when the user visited BioPortal for the first time ever.

3 BiOnIC Schema

To define the BiOnIC RDF dataset schema (shown in **Figure 1**), we used SKOS, the Data Catalog vocabulary (DCAT), and the Provenance Ontology (PROV-O). BiOnIC contains two types of datasets: (1) Class Statistics Datasets, and (2) User Interaction Sequences Datasets. Each dataset contains data for a specific BioPortal ontology.

Class Statistics Datasets. The `bionic:StatDataset` class represents datasets that publish the aggregated statistics for each ontology class. The `bionic:ClassInfo` captures the structural characteristics of the class (e.g., number of subclasses, siblings, depth). The `bionic:ReuseCount` represents the number of ontologies that reuse a specific ontology class and the type of reuse (reuse by IRI or reuse by CUI [4]). The `bionic:RequestCount` represents the total and unique counts of clicks and queries for each ontology class. For additional content, these RDF datasets can be queried in conjunction with the ontologies and ontology mappings in BioPortal [7].

User Interaction Sequences Datasets. The `bionic:SeqDataset` class represents datasets of user interactions sequences for a particular ontology for both the BioPortal WebUI or the BioPortal API (indicated via the `bionic:accessType` property). The anonymized user identifiers are represented as `prov:Agent` instances. We decided to use `prov:Agent` over `prov:Person`, since the underlying IP addresses may indicate an individual user or an organization. A sequence of user interactions—captured as an instance of `bionic:Sequence`—is represented as a list of `bionic:SeqEntity` instances linked via the `bionic:nextEntity` properties.

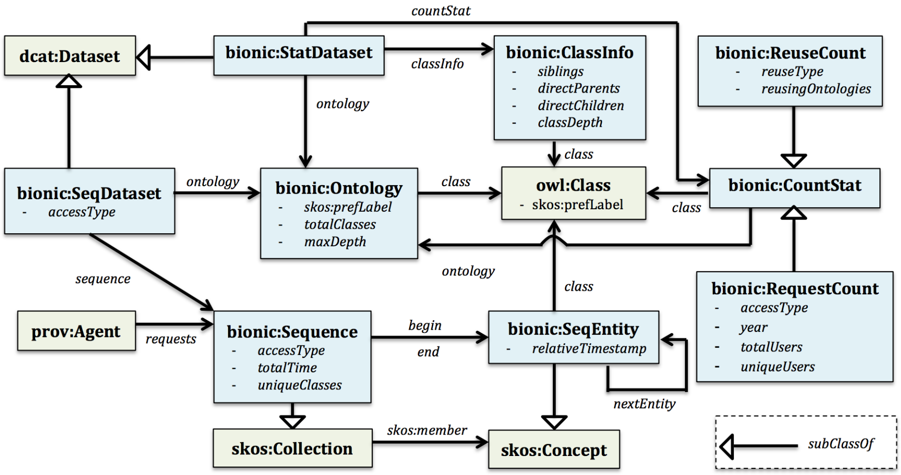


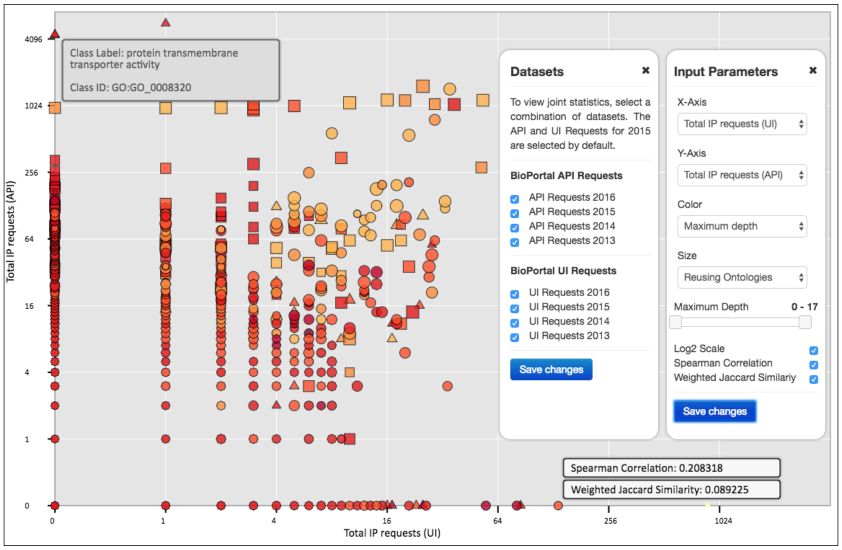
Fig. 1. BiOnIC Schema. BiOnIC facilitates the publishing of two different types of datasets for each ontology in the catalog: *i)* Class Statistics Datasets, which capture the total number of users that explored or queried a particular class, the total number of reusing ontologies, and other class characteristics, and *ii)* User Interaction Sequences Datasets, which capture interaction sequences extracted from BioPortal access logs.

A BiOnIC dataset is associated with one ontology in the BioPortal repository. Users who want to use BiOnIC to perform cross-ontology research (e.g., set of ontologies browsed together, sequence of classes in different ontologies) can easily reconcile these datasets using the `prov:Agent` instances and the `bionic:relativeTimestamp` attributes of the sequence entities. The BiOnIC RDF datasets can be deployed to a SPARQL endpoint, and queries can be formulated in federation with the BioPortal SPARQL endpoint.[3] Example queries include: *i) How many agents click on the subclasses after exploring the parent class in Gene Ontology?*, or *ii) What is the average time spent browsing the Gene Ontology in the BioPortal WebUI?*.

4 VisIOn Web Application

We developed an interactive Web application, VisIOn (Visualizing Ontology Interactions), to help users explore the two types of datasets (class statistics and user interaction sequences) available in the BiOnIC catalog. VisIOn can visualize a dataset in 4 different perspectives: *i)* scatter plot, *ii)* volcano plot, *iii)* word cloud, and *iv)* PolygOnto visualizations [5]. The scatter plot perspective allows the user to select the features to display on the X- and Y-axes, and then visualizes the aggregate statistics along with the structural features of the classes (**Figure 2**).

[3] `http://sparql.bioontology.org/`



Class Label: protein transmembrane transporter activity

Class ID: GO:GO_0008320

Datasets ✖

To view joint statistics, select a combination of datasets. The API and UI Requests for 2015 are selected by default.

BioPortal API Requests

☑ API Requests 2016
☑ API Requests 2015
☑ API Requests 2014
☑ API Requests 2013

BioPortal UI Requests

☑ UI Requests 2016
☑ UI Requests 2015
☑ UI Requests 2014
☑ UI Requests 2013

Save changes

Input Parameters ✖

X-Axis

Total IP requests (UI)

Y-Axis

Total IP requests (API)

Color

Maximum depth

Size

Reusing Ontologies

Maximum Depth **0 - 17**

Log2 Scale ☑
Spearman Correlation ☑
Weighted Jaccard Similariy ☑

Save changes

Spearman Correlation: 0.208318

Weighted Jaccard Similarity: 0.089225

Fig. 2. VisIOn Web Application. An example scatter plot perspective that visualizes the total number of API requests against the WebUI requests for the Gene Ontology [1] between 2013–2016. The user can interactively select the different statistics for visualization, and also use structural features to change the node shape, color and size.

Table 1. Characteristics of the BiOnIC datasets

Feature	WebUI			API		
	Max	Median	Std. Dev.	Max	Median	Std. Dev.
Total Time of Interaction (sec.)	88,644,234	4,738	16,168,783	89,816,694	5	7,087,630
Total Sequence Classes	217,938	3	576	14,062,502	13	42,051
Unique Sequence Classes	52,813	3	153	299,391	2	1,456
% of Classes Accessed	100	22	26	100	18	26
Unique Users	202,163	185	14,888	55,339	153	4,019

The volcano plot perspective is a special kind of scatter plot that visualizes the statistical significance of changes in large datasets. We execute a Fisher's exact test by keeping either the time period or the access mode constant. To correct for multiple hypothesis testing, we use FDR-adjusted (False Discovery Rate) p-values and odds ratios. The Log-10 p-values are plotted against the Log-2 transformed odds ratios for the volcano plot, creating a volcano-shaped scatter plot. The volcano plot helps a user determine if a particular class in a selected ontology is significantly accessed through a particular access mode (e.g., API vs WebUI) or during a particular time period (e.g., 2014 vs 2015). For example, we found that *protein transmembrane transporter activity* and other related classes in the Gene Ontology were significantly accessed during 2015 (Log-10 p-value $\approx$ -7, Log-2 odds ratio $\approx$ -4), compared to 2016.

The word cloud perspective visualizes the labels (`skos:prefLabel`) for the classes that are significantly accessed over different periods of time. The size of these labels is determined from the odd ratios computed from the Fisher's exact test. Finally, in the PolygOnto perspective (**Figure 3**), we visualize the ontology as a graphical polygon shape, and we visualize the different sequences as smaller blue-colored polygons overlaid on the ontology polygon (red polygon). The height of the polygon represents the number of hierarchical layers in the ontology, whereas the width represents the number of classes in each layer. We developed the method to generate the PolygOnto visualization for an ontology and its sequence datasets in our prior work [5].

5 Dataset Characteristics and Availability

BiOnIC publishes the aggregate class statistics and interaction sequences for 255 biomedical ontologies and 515,456 total users. **Table 1** lists the summary statistics (maximum, median, standard deviation) for the user-level features (total and unique entities in a sequence, total time of interaction) and ontology-level features (% of classes accessed, number of unique users) for the two modes of accessing BioPortal (WebUI and API).

The BiOnIC vocabulary, the RDF and the Tab Separated Values (TSV) files for the datasets can be downloaded from: `http://onto-apps.stanford.edu/bionic`. The VisIOn Web application is publicly available at: `http://onto-apps.stanford.edu/vision`. The BiOnIC datasets are also listed at DataHub `https://datahub.io/dataset/bionic`. We plan to update the catalog and DataHub.io listing every 6 months to support the research in the development of intelligent interfaces for ontology development and maintenance, as well as for use in other applications described below.

6 Applications of the BiOnIC datasets

To demonstrate the wide applicability of the BiOnIC catalog, we present briefly in this section three different applications that we have built on top of the BiOnIC datasets. We also discuss other types of applications and research that BiOnIC enables.

6.1 Application 1: Characterizing User Behaviors

BioPortal is a very important resource for the biomedical community. Millions of users have accessed it to drive a wide range of biomedical applications. However, we did not have any insights into how different types of users are accessing BioPortal features. Thus, we used the BiOnIC sequence datasets consisting of both the WebUI and API requests to model the browsing behavior of BioPortal users using memoryless Markov chains [6]. We represented the user behavior as a vector, and we clustered these vectors using *k*-means. We were able to categorize BioPortal users into seven distinct categories (e.g., class explorers, search users). We will use these results to customize the BioPortal user interface to better suit the navigation patterns of the users.

6.2 Application 2: Identifying Exploration and Querying patterns in Ontologies

Ontology reuse is an important guideline in developing ontologies. We investigated if classes that are more heavily accessed through the BioPortal WebUI and API inform the reuse of these classes in other ontologies. For this study, we used the BiOnIC statistics on user clicks, queries and reuse counts for each class in every ontology. We did not find a significant Spearman correlation between class access and reuse.

We also investigated if user browsing behaviors through the BioPortal WebUI and the API correlate with each other. For this reason, we developed the PolygOnto visualization [5] (**Figure 3**) that exploits the class hierarchy to reveal regions in an ontology

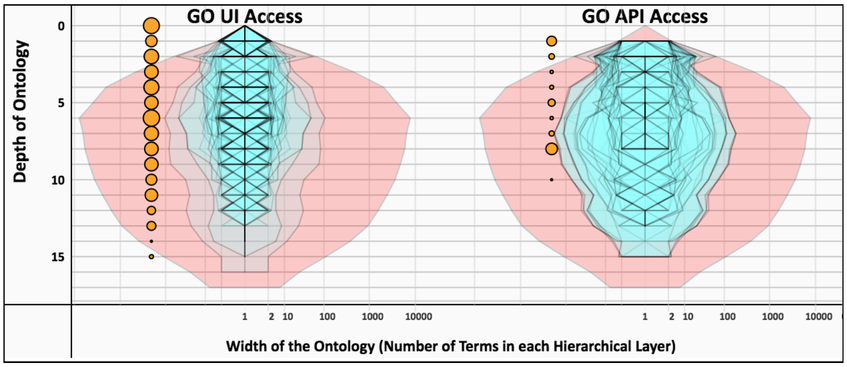


Fig. 3. PolygOnto Visualization. This example PolygOnto visualization displays the user interactions with the Gene Ontology [1] using through the BioPortal WebUI and the API. The underlying red polygon depicts the class hierarchy, whereas each smaller blue polygon represents the set of ontology classes accessed. The width represents the number of classes at a given depth in the hierarchy. Lower levels of the ontological hierarchy are rarely explored or queried.

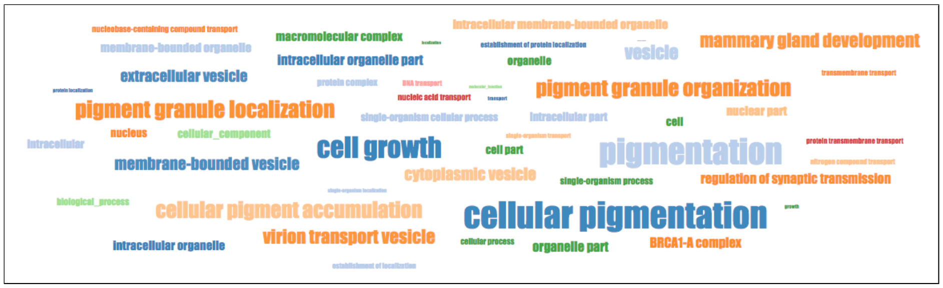


Fig. 4. Word Cloud Perspective. Gene Ontology classes significantly accessed using the BioPortal API in 2015, when compared to 2016.

where users tend to explore and query more. Classes are arranged in different layers based on the maximum depth from the root class in the ontology. The height of a polygon is indicative of the number of layers, and the width at each layer is proportional to the number of classes in that layer. The underlying red polygon depicts the ontological hierarchy. The smaller blue polygons represent the set of ontology classes accessed by a user. We observe two types of exploration patterns: *i) Triangles:* 1 parent → 2 child classes, and *ii) Inverted Triangles:* 1 child → 2 parent classes. We also observe that classes in the lower levels of the class hierarchy are rarely explored or queried by users.

6.3 Application 3: Comparing BioPortal Access Modes and Temporal Influences

Understanding how users access different parts of an ontology in BioPortal (through WebUI or API) can reveal trends into the evolution of an ontology. We use Fisher's exact test over the BiOnIC aggregative statistics, and multiple hypotheses testing, to investigate if users access certain ontology classes significantly more when compared between different access modes and different time periods. The VisIOn Web application (see Section 4) provides fascinating insights in the influence of access modes and time on information retrieval in ontologies. For example, as seen in the word cloud perspective of VisIOn (**Figure 4**), more users queried the Gene Ontology using the BioPortal API for classes related to *pigmentation* in 2015, when compared to 2016. **Figure 2** shows that certain classes (e.g., *protein transmembrane transporter activity*) are requested multiple times using the BioPortal API, but are never requested using the BioPortal WebUI. Moreover, by observing the VisIOn word cloud and the volcano plot perspectives, we can observe the rise of queries for certain classes related to *Zika virus* and *Ebolavirus* in several disease ontologies. These discrepancies may be due to temporal influences or other reasons, which can be ascertained by domain experts.

6.4 Applications and Areas of Research Enabled by BiOnIC

The BiOnIC catalog can enable further applications and areas of research. For example, our study to categorize user browsing behaviors can be extended to incorporate the structural features of the ontology classes. The insights gathered from such a study will support the development of personalized user interfaces for ontology navigation, which will take into account the user type and the predictions of the next class that a user is likely to access. These insights may be generalizable to other ontology repositories that feature similar interfaces for user browsing. Ontology summarization and modularization are active research areas that aim to reduce the size and complexity of large ontologies to enable users to better understand ontologies and to use ontologies more efficiently in downstream applications. The BiOnIC datasets may be used as features in developing advanced methods for ontology summarization and modularization.

7 Conclusion

The BiOnIC catalog publishes aggregate class statistics and sequences of user interactions with biomedical ontologies as observed in the BioPortal repository. Using the best practices for anonymization, vocabulary reuse and publishing Linked Data, we made these datasets available as RDF files. We also presented the VisIOn Web-based visualization application that offers different perspectives of these datasets to biomedical researchers, practitioners and ontology developers. We showcase three applications that are built on top of the BiOnIC catalog to demonstrate how this resource will be valuable to ontology developers, repository maintainers and domain users. To the best of our knowledge, this is the first attempt to publish user interactions data from a widely-used semantic application. Semantic Web researchers can analyze these datasets to gain more insights into the interaction and behavior of users when browsing, querying and reusing ontologies. We envision, that the analysis of the BiOnIC datasets will spur the development of a variety of novel applications, such as, intelligent and intuitive interfaces for ontology browsing and editing.

Acknowledgments

This work is supported in part by grants U54-HG004028 and GM086587 from the US National Institutes of Health.

References

1. Ashburner, M., et al.: Gene Ontology: tool for the unification of biology. Nature genetics 25(1), 25–29 (2000), dOI:10.1038/75556
2. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of medical informatics p. 67 (2008)
3. Housley, R.: A 224-bit one-way hash function: Sha-224 (2004)
4. Kamdar, M.R., et al.: A systematic analysis of term reuse and term overlap across biomedical ontologies. Semantic Web (Preprint), 1–19 (2016)
5. Kamdar, M.R., et al.: Analyzing user interactions with biomedical ontologies: A visual perspective. Journal of Web Semantics (under review) (2017), `https://goo.gl/qmQBLE`
6. Walk, S., et al.: How users explore ontologies on the web: A study of NCBO's bioportal usage logs. In: Proceedings of the 26th International Conference on World Wide Web. pp. 775–784. WWW '17 (2017)
7. Whetzel, P.L., 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), W541–W545 (2011)