How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs

Simon Walk
Stanford University
walk@stanford.edu

Lisette Espín-Noboa
GESIS & University of Koblenz-Landau
lisette.espin@gesis.org

Denis Helic
Graz University of Technology
dhelic@tugraz.at

Markus Strohmaier
GESIS & University of Koblenz-Landau
markus.strohmaier@gesis.org

Mark A. Musen
Stanford University
musen@stanford.edu

ABSTRACT

Ontologies in the biomedical domain are numerous, highly specialized and very expensive to develop. Thus, a crucial prerequisite for ontology adoption and reuse is effective support for exploring and finding *existing* ontologies. Towards that goal, the National Center for Biomedical Ontology (NCBO) has developed BioPortal—an online repository containing more than 500 biomedical ontologies. In 2016, BioPortal represents one of the largest portals for exploration of semantic biomedical vocabularies and terminologies, which is used by many researchers and practitioners. While usage of this portal is high, we know very little about how exactly users search and explore ontologies and what kind of usage patterns or user groups exist in the first place. Deeper insights into user behavior on such portals can provide valuable information to devise strategies for a better support of users in exploring and finding existing ontologies, and thereby enable better ontology reuse. To that end, we study and group users according to their browsing behavior on BioPortal and use data mining techniques to characterize and compare exploration strategies across ontologies. In particular, we were able to identify seven distinct browsing types, all relying on different functionality provided by BioPortal. For example, *Search Explorers* extensively use the search functionality while *Ontology Tree Explorers* mainly rely on the class hierarchy for exploring ontologies. Further, we show that specific characteristics of ontologies influence the way users explore and interact with the website. Our results may guide the development of more user-oriented systems for ontology exploration on the Web.

Keywords

Browsing behavior; Semantic Web; BioPortal; Markov chain; Stationary distribution; Clustering

WWW 2017, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052606

1. MOTIVATION

Facilitating reuse on the Semantic Web requires support for effectively finding and exploring existing semantic resources such as ontologies or taxonomies. Particularly in the biomedical domain, where ontologies are highly specialized, costly and often created collaboratively by large groups of domain experts, *identifying and finding existing semantic vocabularies and terminologies* is crucial. To support researchers and practitioners in this task, the National Center for Biomedical Ontology (NCBO) has developed BioPortal[1]—*the world's most comprehensive online repository of biomedical ontologies* [21, 16, 33, 18].

Problem. In this paper, we want to shed light on how users are exploring ontologies and terminologies in BioPortal. Without a deep understanding of how to facilitate the exploration of existing ontological resources on the Web, increasing the reuse of ontologies on the Semantic Web will remain an elusive goal. Towards that goal, we want to (i) cluster traces of user interactions to obtain different exploration behavior types and (ii) investigate ontologies in terms of their user behavior. New insights into user behavior on such portals can potentially provide actionable information to devise strategies that better support researchers and practitioners in exploring existing ontologies, and thereby may enable better reuse of ontological content in general.

Approach. To identify different types of user behavior on BioPortal, we adopt a clustering approach based on (i) dynamic browsing features extracted from the Apache logs of BioPortal and on (ii) calculating stationary distributions from first order Markov chain representations of user transitions. We demonstrate the effectiveness of our approach and investigate if and to what extent exploration strategies of users differ when exploring biomedical ontologies on BioPortal. In addition, we use Principal Component Analysis (PCA) to visually inspect the obtained clusters.

Contributions. We present an approach for identifying different user behavior types in ontology exploration log data using stationary distributions from first order Markov chains. We provide evidence for a total of seven distinct browsing-behavior types on BioPortal and offer an interpretation of their relevance and meaning. We use these different types of exploration behavior to characterize ontologies in terms of the different user interactions they attract. We

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

Table 1: Overview of all actions available in the BioPortal user interface.

	Action Labels	Description
Main Page	*Browse Main Page*, *Browse Ontologies*, *Browse Search*, *Browse Help*, *Browse Mappings*, *Browse Recommender*, *Browse Annotator*, *Browse Resource Index*, *Browse Projects*, *Browse Notes*	Browsing main areas and functionality of BioPortal.
Ontology Page	*Ontology Summary*, *Browse Ontology Classes*, *Browse Ontology Class*, *Browse Ontology Class Tree*, *Browse Ontology Mappings*, *Ontology Analytics*, *Browse Ontology Widgets*, *Browse Ontology Visualization*, *Browse Ontology Notes*, *Browse Ontology Properties*, *Browse Widgets*, *Browse Ontology Property Tree*, *Browse Class Notes*	Labels for actions that can be performed while browsing a specific ontology on BioPortal.
Edit Content	*Create Ontology Submission*, *Validate Ontology File*, *Virtual Appliance Download*, *Browse Ontology Submission*	Actions triggered when uploading new versions of ontologies.
User Account	*Login*, *Log-Out*, *Sign-Up*, *Lost Password*, *Browse Account*, *Feedback*	Actions to manage accounts on BioPortal.
Control Term	*BREAK*	$\geq$30 minutes inactivity between two actions.

find that particular characteristics of semantic resources, as represented on BioPortal, influence how users interact with them. For example, BioPortal lacks the ability to visualize flat ontologies (i.e., ontologies without a structured hierarchy), which fuels the emergence of diverse exploration strategies as identified by our approach. Overall, our work advances our understanding of ontology exploration behavior on prominent repositories of biomedical ontologies on the Web.

2. RELATED WORK

Biomedical Ontology Repositories. There exists a vast variety of repositories for ontologies and taxonomies with similarly diverse core interests. We provide a brief overview of the most important projects from the biomedical domain and describe how they are related to BioPortal. The Open Biomedical Ontologies (OBO) Foundry [25] initiative collects and maintains a set of different ontologies, which are specifically designed for interoperability. The whole library of ontologies is available on GitHub[2] and developers of ontologies aspire to have their content included in the OBO Foundry. All ontologies of the OBO Library are automatically imported into BioPortal.

The European Bioinformatics Institute [11] maintains the Ontology Lookup Service which provides an alternative mechanism for accessing content in the OBO Library.

The Unified Medical Language System (UMLS) [1] is a collection of biomedical terminologies and ontologies distributed by the National Library of Medicine. Terms in UMLS are mapped to one another through the UMLS Metathesaurus. Similarly to OBO Foundry, the content of UMLS is also available for exploration in BioPortal.

The majority of research about (biomedical) ontology repositories focuses on the selection and maintenance (including provenance) of the included ontologies and taxonomies. In this paper, we present analyses that complement existing research by studying aspects about the interaction behavior of users on BioPortal [21, 16, 33, 18]—one of the largest biomedical ontology and taxonomy repositories on the Web.

User Interactions on the (Semantic) Web. A lot of research has addressed the tasks, actions, tools and processes required for developing and evaluating ontologies [8]. This has led the Semantic Web community to develop a variety of different guidelines and methodologies [23] or compile best practices [17] for engineering ontologies.

Engineering ontologies collaboratively can potentially lead to structural conflicts since several users can edit ontologies simultaneously. To circumvent this, Hellman and Gal [9] propose locking mechanisms based on a graph depiction of the concept dependencies. Pesquita et al. [19] leveraged the location and specific structural features to show that these can be used to determine if and where the next change is going to take place in the Gene Ontology[3].

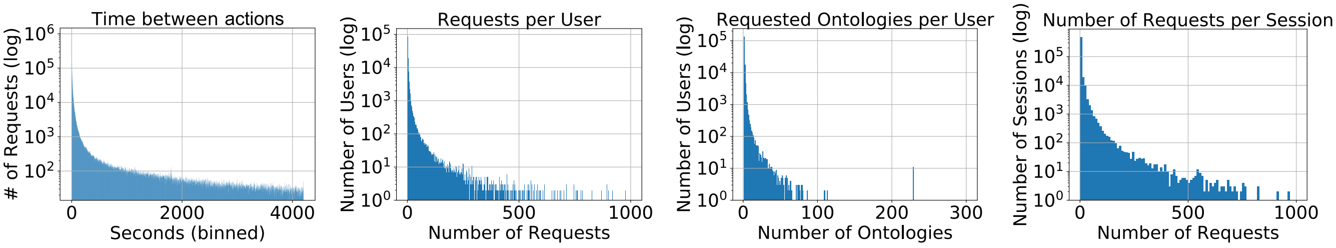
To learn more about the impact and quality of collaboration in collaborative ontology-engineering projects, Strohmaier et al. [26] investigated the hidden social dynamics that take place in such projects from the biomedical domain and provided new metrics to quantify various aspects of the collaborative engineering processes. User roles have been also studied from many different angles. For example, Falconer et al. [6] investigated and classified users according to different roles in collaborative ontology-engineering projects. In 2014, Van Laere et al. [27] used K-means and the GOSPL methodology to classify users by clustering interactions that users engage in, while engineering an ontology. Wang et al. [32] used association-rule mining on the change-logs of collaborative ontology-engineering projects, and showcased the utility of the identified editing patterns in a prediction experiment. Walk et al. [28] studied user editing trails of ontology-engineering projects by defining and comparing a set of hypotheses about how users edit ontologies.

Further, there has been increasing interest on analyzing and categorizing human sequences in the past, including mobility [5], page access [7, 12], edit [28, 30, 31] as well as click stream logs [24]. Additionally, Markov models of various orders have been employed [2, 15, 20, 4, 22, 29] to model and predict clicks or interactions of users on the (Semantic) Web. Recently, Chierichetti et al. [3] conducted an analysis that questioned if a first-order model best represents the navigation behavior of humans on the Web. In this direction, the work in Lakshimarayan et al. [13] demonstrates that higher order Markov chains can be used to model Web browsing behavior and predict user intent towards buying specific products.

Making sense and interpreting such sequential data is not a trivial task. The work in Hoxha et al. [10] proposes a platform to analyze user browsing patterns by semantically formalizing the usage logs of Web data, providing means to query and retrieve sessions of users that satisfy certain semantic and temporal conditions. Research has also been

[2] http://obofoundry.org

[3] http://www.geneontology.org



(a) Seconds between Requests (b) Requests per IP (c) Ontologies per User (d) Requests per Session

Figure 1: *BioPortal Characteristics.* Entities on the x-axes for all figures are binned according to the displayed values and cut-off at their maximum values for reasons of readability. The seconds between requests (a) are heavy-tailed, meaning that the majority of requests are conducted within short periods of times. The number of request actions per IP (b), a proxy for users on BioPortal, is heavy-tailed. The majority of users have visited only a total of one or two ontologies and (c) the majority of sessions exhibit a small number of requests (d).

focused on human behavior beyond browsing patterns and study reading behavior of users in Wikipedia as well as the stability of the reading patterns per user [14]. A more generic approach is proposed in [24] where the authors present HypTrails—a Bayesian methodology that allows researchers to compare and rank hypotheses about digital trails on the Web, which has also been applied on the edit logs of collaborative ontology-engineering projects by Walk et al. [28].

To the best of our knowledge, analyzing how users explore ontologies in the context of ontology repositories on the Web has not been addressed as in-depth before.

3. MATERIALS & METHODS

3.1 BioPortal

BioPortal—an online biomedical ontology and taxonomy repository—was created by the National Center for Biomedical Ontology (NCBO). The main goals of NCBO involve not only the creation and maintenance of a comprehensive repository of biomedical ontologies and terminologies, but also the task of building novel tools and Web services to enable and augment the (re)use of all stored semantic vocabularies and terminologies in clinical and translational research. BioPortal currently contains more than 500 biomedical ontologies, and supports a wide range of Web services, such as using ontologies for annotating resources and generating value sets or specific ontology views. Additionally, BioPortal allows users to explore ontological content not only by using a standard tree browser, but it can also visualize resources using custom tailored widgets, which help users in comprehending the complexity of large biomedical resources (see Table 1 for all click types). Similarly, the website provides functionality to explore mappings between ontologies, which can be used to directly compare the use of related terms as well as the overlap between different ontologies.

3.2 Data Acquisition & Preprocessing

For the analyses presented in this paper, we have parsed all the requests stored in the Apache logs of BioPortal from 12/31/2015 to 07/19/2016 (around 7 months; cf. Table 2). Each line in the request logs contains information (among others) about the IP, the timestamp, the actual resource that was requested and the useragent, which is used to provide additional information about the requester, such as browser, operating system or name of the bot/spider, for each request.

Preprocessing. In a first step, we reduced a total of roughly 51.8M entries (7 months) to 16.7M by removing all requests that were conducted by bots, spiders and other automatic scripts via the combination of existing useragent and IP blacklists as well as through manual inspections.

Then, we reduced our dataset to only include requests, which were triggered by users and involve an actual interaction with the website. For example, if users click on an ontology name, many automatic requests (AJAX calls) are triggered, which do not represent interactions of users, and are hence filtered from our dataset. Note that we consider all interactions that trigger a request for our analyses. For example, we consider search queries entered by users on BioPortal as interactions with the website. After preprocessing, we still have a total of 2.52M requests (see Table 1), generated by 215,908 unique IP addresses over the course of 7 months. Note that the number of different ontologies that users requested is larger than the number of stored ontologies in BioPortal. Whenever users manually type in requests in the browser, typos can occur. Further, ontologies on BioPortal are also not only accessible through their abbreviations but also through their internal IDs. In both situations, the number of unique ontologies in the request logs increases, while the number of stored ontologies remains unaffected.

Dataset Characteristics. The distribution of time between requests (seconds) is depicted in Figure 1(a). The seconds between requests are heavy-tailed, meaning that the majority of requests are performed within 0-10 seconds.

To be able to asses if and to what extent we are able to observe power users—very active users who contribute the

Table 2: Characteristics of the BioPortal dataset.

Feature	Value
Unique IPs	215,908
Unique IPs (with $\geq$ 2 requests)	168,008
Ontologies	1,818 (517)
Actions (ca.)	2.52M
Sessions (time between requests $\geq$ 30 mins)	513,659
1-request sessions	165,543
# requests in $\geq$ 2-request sessions	2,36M
Average/Median session duration	217s/0s
First request	2015/12/31
Last request	2016/07/19
Observation period (ca.)	7 months

Table 3: *Example of a session.* Using the Apache logs of BioPortal, we can chronologically order the requests of each user (identified by IP) individually, assign action labels to each request and create a sequence of actions, which can not only be used to fit first-order Markov chains, but also to calculate stationary distributions for each user.

Timestamp	Type	Request	Action Labels (Sequence Step)
2016-03-14 09:07:32	GET	/	Browse Main Page (1)
2016-03-14 09:07:46	GET	/login?redirect=http%3A%2F%2Fbioportal.bioontology.org%2F	Login (2)
2016-03-14 09:07:48	POST	/login	Login (3)
2016-03-14 09:07:50	GET	/	Browse Main Page (4)
2016-03-14 09:08:04	GET	/ontologies/MCCV	Ontology Summary (5)
2016-03-14 09:08:22	GET	/ontologies/MCCV/submissions/new	Create Ontology Submission (6)
2016-03-14 09:09:34	POST	/ontologies/MCCV/submissions	Create Ontology Submission (7)
2016-03-14 09:09:59	GET	/ontologies/success/MCCV	Create Ontology Submission (8)
2016-03-14 09:10:14	GET	/ontologies/MCCV	Ontology Summary (9)

Sequence: Browse Main Page –> Login –> Login –> Browse Main Page –> Ontology Summary –> Create Ontology Submission –> Create Ontology Submission –> Create Ontology Submission –> Ontology Summary

majority of all requests—we have grouped users (in the form of IP addresses) according to the number of requests they have contributed and visualized the results in Figure 1(b). While the majority only conduct a very small number of requests (between 1-10), we were able to identify 30 users who conducted more than $5,000$ interactions on BioPortal within 7 months. Additionally, we were able to identify that the majority of users only visit one or two ontologies on the website (cf. Figure 1(c)).

Finally, we aggregated all requests of users into sessions. Each session contains all chronologically sorted interactions of one user, unless two interactions are apart more than 30 minutes. In that case, the current session is closed, a new session is opened and the process is repeated. We have selected 30 minutes as threshold, as the requests per IP are heavy-tailed and the majority of timespans between requests are within that threshold, introducing only a small number of sessions. We were able to group all of the 2.52M requests into a total of $513,659$ sessions, where $165,543$ only consist of a single interaction. As depicted in Figure 1(d), the majority of sessions in BioPortal exhibit a low number of requests, while only very few sessions last for more than 100 requests.

Action Label Mapping. Finally, we created a set of regular expressions to map each request to easier-readable labels (cf. Table 1 for all the different types of labels). In general, the generated action labels can be grouped into 5 different categories. The two most important groups of action labels are the ones accessible from the main page of BioPortal, such as *Browse Ontologies* or *Browse Search*, and the ones that describe actions conducted on single ontology pages, such as *Browse Ontology Class* or *Browse Ontology Property Tree.*

Action Sequence Generation. Using the action labels we are able to extract easily readable chronological sequences of actions for each user on BioPortal (see Table 3). Whenever an IP address exhibits two or more sessions, we connect the sessions with a *BREAK* state to be able to see where users stop and resume browsing BioPortal.

3.3 Modeling Browsing Behavior

We model user browsing behavior on BioPortal with memoryless Markov chains. A Markov chain consists of n states s_i from a finite state-space S with $n = |S|$ and a transition matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ where p_{ij} defines the probability to traverse from state s_i to state s_j. Since each row in $\boldsymbol{P}$ defines a probability distribution for each i we have $\sum_j p_{ij} = 1$.

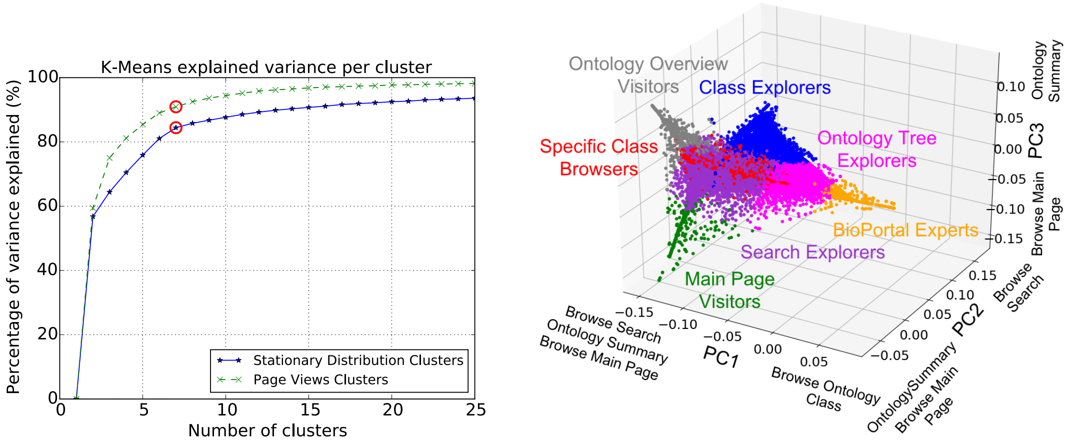
We represent each action on BioPortal as a single state of a Markov chain (see Table 1). Then, an element p_{ij} from the transition matrix $\boldsymbol{P}$ represents the probability of performing action j after action i has been performed.

To compare users with each other we compute $\boldsymbol{P}$ and its stationary distribution for each individual user. In our case, the stationary distribution is a probability distribution over actions, which defines how likely we will find a user performing a given action in the limit of large number of steps. Generally, the stationary distribution captures the way how users move between actions, i.e. it encodes the sequential interactions within an n-dimensional vector, and therefore provides additional information over simple view counts (cf. Table 4).

To compute stationary distributions, we first define the weighted matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, where each element a_{ij} is set to the number of observed transitions between states i and j in our empirical dataset (see Table 1 for all possible states). A

Table 4: *Stationary Distribution and Page Views Illustration.* Given a vector of page views (2xA, 2xB and 2xC) the stationary distribution depends on the order of the appearance of each state in the sequence. For example, given a sequence of "ABCABC", we create a weighted matrix by counting the transitions between the three states A, B and C. Normalizing each row produces the transition matrix *P*, which we use to calculate the stationary distribution *π* (top row). For a user with the same page views, but a different sequence, we obtain a different stationary distribution (bottom row). The static page views remain unaffected by the ordering of the sequence, as only the total number of occurrences of each state is considered.

Sequence	Weighted Matrix	Transition Matrix	Stationary Distribution	Page Views
ABCABC	$\boldsymbol{A} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 1 & 0 & 0 \end{pmatrix}$	$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$	$\boldsymbol{\pi} = \begin{pmatrix} 0.3533 \\ 0.3356 \\ 0.3111 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$
AABBCC	$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}$	$\boldsymbol{\pi} = \begin{pmatrix} 0.5229 \\ 0.2916 \\ 0.1855 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$



(a) Selection of K

(b) Stationary Distribution Clusters

Figure 2: *User Behavior Clusters.* To estimate the number of clusters to investigate, we have plotted the percentage of the explained variance (y-axis) per cluster (x-axis) in (a) for the stationary distribution clusters (solid blue) and the page view clusters (dashed green). We select the number of clusters (7) to a value where the introduction of new clusters only minimally increases the explained variance (red circles). To further investigate the results of K-means, we have reduced and visualized the stationary distribution vectors of each user to three dimensions using principal component analysis in (b). Each point represents one user of our dataset, while the colors represent the corresponding clusters obtained by K-means (before dimensionality reduction). We provide labels for the clusters and the extremes of the axes. The latter represent the actions with the smallest and highest coefficients of the corresponding principal components (see Table 5).

transition matrix $\boldsymbol{P}$ has to satisfy certain conditions to have a stationary distribution, most notably irreducibility (each action has to be reachable from all other actions) and aperiodicity (the return times to actions have to be aperiodic).

To guarantee both irreducibility and aperiodicity, we add a teleportation factor α to $\boldsymbol{A}$, which (technically) would allow users to teleport between all actions with a very small probability, similar to PageRank. The teleportation factor connects each state to all others (satisfying irreducibility) including a self-loop (satisfying aperiodicity):

$$\boldsymbol{W} = \boldsymbol{A} + \frac{\alpha}{n}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}, \tag{1}$$

where we set $\alpha = 0.15$ and $\mathbf{1}$ is a vector of all ones from $\mathbb{R}^n$. Finally, we normalize each row in $\boldsymbol{W}$ to sum up to 1 to obtain the transition matrix $\boldsymbol{P}$.

Now, the stationary distribution $\boldsymbol{\pi}$ can be calculated as the left eigenvector of $\boldsymbol{P}$ with its largest eigenvalue 1 and the stationary distribution satisfies the eigenvalue equation for the matrix $\boldsymbol{P}$: $\boldsymbol{\pi}^{\boldsymbol{T}} = \boldsymbol{\pi}^{\boldsymbol{T}}\boldsymbol{P}$. An example calculation of the stationary distribution (without the teleportation factor) is depicted in Table 4. Assuming two users exhibit the same number of page views on the pages A, B and C, we can incorporate the information about the chronological ordering of the sequence by calculating the stationary distribution. Hence, a sequence of “ABCABC” page visits (top row of Table 4) yields a different stationary distribution than a sequence of “AABBCC” page visits (bottom row of Table 4).

3.4 Clustering Browsing Behavior

To deepen our understanding about the browsing behavior of users on BioPortal, we first group similar users. To that end, we use K-means—an unsupervised clustering algorithm—which clusters “close-by” users. For K-means we need to embed users in a vector space so that their proximity can be determined by, for example, pairwise Euclidean distances between user vectors. For that purpose we represent users with their stationary distribution vectors.

Once, when users are represented as vectors, K-means randomly selects K initial central vectors (centroids), where K has to be manually set and defines the number of clusters the algorithm will find. In every iteration, each user vector is assigned to the closest centroid. The positions of the centroids are then updated (i.e., moved towards the center of the corresponding vectors).

This process continues until convergence, that is until the displacement of the centroids between iterations is below a certain threshold. Note that the choice of the clustering algorithm is arbitrary and we leave it open to future work, to determine which algorithm is best suited for a specific application. For illustration purposes, we have chosen K-means as it is well studied and can handle large datasets (168,008 distinct users, clustered over 34 action-label dimensions).

4. RESULTS

4.1 Differences in User Browsing Behavior

We have calculated the stationary distribution for each user as described in Section 3.4, and used K-means to group them accordingly. Additionally, we clustered users using static page view vectors.

Estimating & Validating *K*. To estimate a plausible number of clusters, which best represent the different types of browsing behavior in BioPortal, we have calculated the percentage of variance explained by 1 to 25 clusters for our empirical data (cf. Figure 2(a)). This method is also often referred to as “elbow”-method, and states that the optimal value for K, given empirical data, should be set so that the introduction of additional clusters only minimally increases

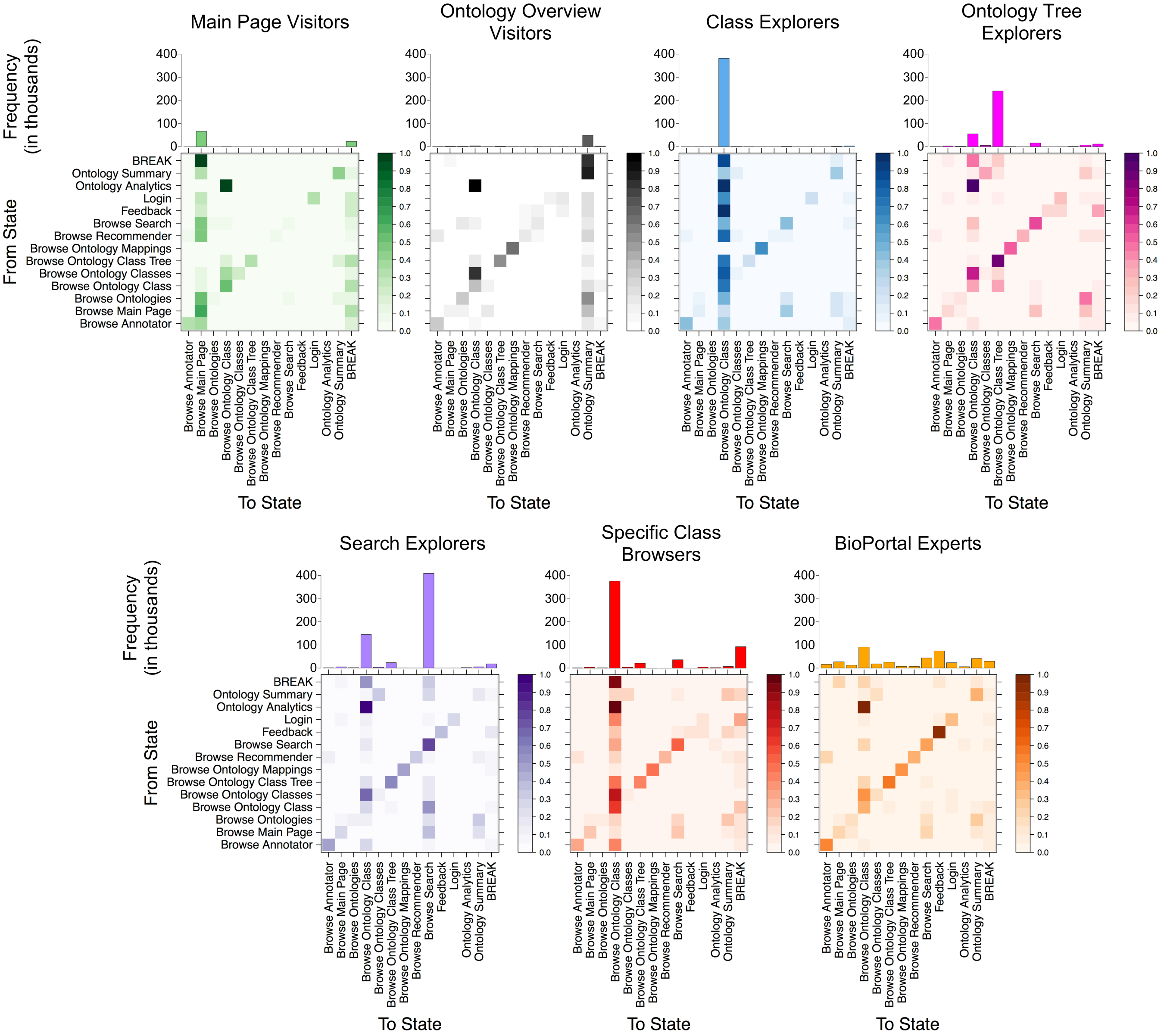


Figure 3: *Browsing-Behavior Types*: We have extracted and visualized the transition probabilities between the different actions for all 7 clusters obtained by K-means when using the stationary distribution vectors (bottom). Transitions are always read *From State* (left) to *To State* (bottom) and rows are normalized individually. The darker the colors, the higher the transition probability between two states. Histograms of the actions are depicted on top of the transition matrices and indicate their absolute occurrences in the extracted sequences for each cluster. All of the clusters exhibit differences in the frequencies (histograms) and transition probabilities.

the explained variance. For both our experiments, the clustering via the stationary distributions and page views, the best value for K is 7.

To be able to visually inspect and validate the resulting clusters, we have extracted three principle components by applying PCA on the stationary distribution vectors and visualized the results (see Figure 2(b)). Each point represents one user of our dataset and the colors indicate one of the seven clusters suggested by K-means. Additionally, we have added (i) the action labels with the largest and smallest coefficients (see Table 5) to the corresponding axes to allow for manual interpretation and (ii) provide labels for the different clusters. We can see that the clusters obtained by K-means, when using the stationary distribution vectors, are easily distinguishable and warrant further inspection.

Categorizing Clusters. Hence, we have extracted and visualized the action sequences of all clusters (see Figure 3; the colors correspond to the clusters in Figure 2(b)) individually. Note that we have removed actions with very few occurrences from the visualizations for reasons of readability, which is why the transition probabilities of the rows in Figure 3 do not necessarily sum up to 1.0.

Using the stationary distribution as input for K-means, we were able to obtain a total of 7 distinguishable browsing-behavior types on BioPortal:

The **Main Page Visitors** primarily visit the main page of BioPortal and exhibit very few other actions. The cluster consists of 2,813 users with an average of 32.2 actions (median of 2) per user.

The **Ontology Overview Visitors** consist of 2,668 users, who primarily visit ontology overview pages and refrain from exploring the classes of the ontologies. The average user in this group performs 22.7 actions (median of 3).

The group of **Class Explorers**, the biggest cluster with a total of 109,159 users, specifically target and visit classes of the different ontologies hosted on BioPortal. The average number of actions for this group is 3.6 (a median of 2).

Ontology Tree Explorers use the hierarchical representation of an ontology on BioPortal to explore and browse the content of an ontology. A total of 4,490 users belong to this group with an average of 77.3 actions (median of 14).

A total of 6,701 **Search Explorers** make extensive use of the search functionality provided by BioPortal to identify, explore and find classes in the ontologies. On average, users of this cluster conduct 92.1 actions (median of 18).

Specific Class Browsers represent the second largest browsing behavior type with 22,436 users. In contrast to *Search Explorers*, this group exhibits shorter browsing sessions, evident in the higher number of *BREAK* states, and concentrates on exploring multiple classes of a specific area of an ontology (as opposed to whole ontologies, by inspecting the ontology class tree). The average number of actions for users in this group is 24.9 with a median of 8.

The final group are **BioPortal Experts**, who use several features that BioPortal provides to find and explore ontologies and classes, such as the annotator and recommender. On average, the 19,741 users of this cluster performed 24 actions (median of 4).

Comparison to Page Views. We repeated the same experiment with the static page view vectors as input for K-means and manually inspected the results for $K = 7$ clusters. In contrast to the results shown in Figure 2(b), we were only able to distinguish a total of four different kinds of browsing-behavior types. The biggest group of users are the *Search Explorers*, consisting of a total of 167,979 users and a total of four different clusters with 167,631, 1, 21 and 326 users respectively. We consider them as single cluster as the actions and transition matrices of these groups were nearly indistinguishable upon manual inspection. The three remaining clusters, *Class Explorers*, *Ontology Tree Explorers* and *Feedback Providers*, are again very small, with only 3, 6 and 20 users respectively.

Increasing K. Additionally, we have manually explored up to 14 clusters obtained from K-means, using the stationary distributions as input. Due to the information inherent in the stationary distribution, we could further refine the obtained class labels. In additional to the seven clusters outlined before, we identified *Annotators* (542 users), *Widget Browsers* (399 users), *Search & Tree Explorers* (3,685 users), *Feedback Providers* (9,311 users) and *Account Maintainers* (374 users).

Table 5: *Principal Component Coefficients.* The explained variance for the first principal component (PC1) is 63.67%, **for the second (PC2)** 73.57% **and** 80.53% **for the third (PC3).**

Action Label	PC1	PC2	PC3
Browse Search	−0.1897	+0.9088	−0.0368
Ontology Summary	−0.1253	−0.2736	+0.6240
Browse Main Page	−0.1236	−0.2785	−0.7712
Browse Ontology Class	+0.9571	−0.0702	−0.0127

4.2 Differences when Exploring Ontologies

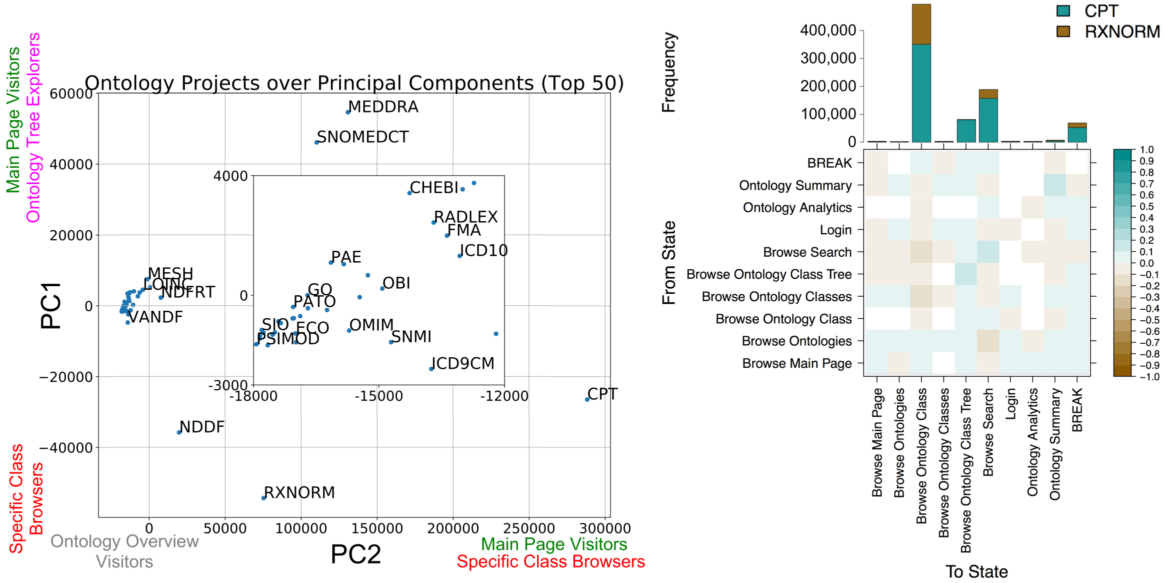
To be able to learn more about exploration and usage strategies of users, we were interested in identifying differences in the distribution of browsing-behavior types that interact with the different ontologies on BioPortal. Hence, we have determined the number of users for each browsing-behavior type cluster for the 50 most visited ontologies in our dataset and aggregated the corresponding number of actions. Finally, we have applied PCA on the aggregated numbers of requests per cluster and visualized the results (similar to Figure 2(b)) to identify differences in the browsing-behavior types between projects (see Figure 4(a)). The larger the difference in actions performed by different browsing-behavior clusters for each project, the larger the distance between the projects.

Due to limitations in space, we restrict the presentation of results to two projects. In particular, we compare the browsing behavior of users for the *Current Procedural Terminology* (CPT) as well as *RxNorm* (RXNORM). The former is a terminology, which is maintained by the American Medical Association, and is used to describe services in the medical, surgical, and diagnostic domain (mostly exclusive) for billing purposes in the US. The latter was developed by the National Library of Medicine as a public resource for various applications, and represents a terminology that contains all medications available on the US market. We have selected CPT and RXNORM, as they are among the ontologies that have received the most visits, exhibit substantially different rankings for the actions per browsing-behavior cluster and were initially designed for different practical applications.

Sequence Extraction. For this analysis, we have extracted all sequences of all users, who performed at least 20% of all their actions on one of the corresponding projects. As a consequence, sequences are not mutually exclusive, meaning that one user could be present in both of our extracted action sequences if that user performed at least 20% of all clicks on CPT and another 20% on RXNORM. Given that the majority of users only visit one or two ontologies (see

Table 6: *Actions for CPT & RXNORM.*

User Type	**CPT** Actions	**RXNORM** Actions
Class Explorers	120,926 (3)	49,440 (2)
Specific Class Browsers	187,987 (2)	88,068 (1)
BioPortal Experts	19,934 (5)	12,755 (4)
Main Page Visitors	195,030 (1)	35,869 (3)
Ontology Tree Explorers	78,485 (4)	1,478 (5)
Ontology Overview Visitors	252 (7)	69 (7)
Search Explorers	1,142 (6)	224 (6)



(a) Ontology Projects over Principal Components

(b) Comparison of Exploration Behavior (CPT & RXNORM)

Figure 4: *Comparison of Exploration Behavior.* We have aggregated the number of actions per browsing-behavior cluster, applied PCA and plotted the landscape of exploration behavior types for the 50 most visited ontologies (blue points) in (a). The extremes of the axes correspond to the clusters with the largest and smallest coefficients of PCA respectively. We have added labels for the ontologies and terminologies where appropriate. The inset magnifies the densely concentrated group of ontologies. The larger the difference in actions performed by different browsing-behavior clusters for each ontology, the larger the distance between ontologies. The detailed comparison of CPT (green) and RXNORM (brown) is depicted in (b). The histogram (top) depicts how often the different actions occurred in the extracted sequences in both ontologies. The transition matrix (bottom) depicts the importance of the transitions between actions of users while browsing CPT (blue) and RXNORM (brown). Transitions that are equally important in both projects are white. The transition from *Browse Search* (From State) to *Browse Ontology Class* (To State) is more dominant for RXNORM. Additionally, RXNORM exhibits a very small number of *Browse Ontology Class Tree* actions.

Figure 1(c)) and that there are many actions, which can not be assigned to an ontology, 20% of all actions already represents a very high filtering criterium. Overall, we collected a total of 61, 367 unique IP addresses for CPT, which conducted on average 10.7 actions (median of 2). In contrast, we extracted action sequences of 23, 757 IP addresses for RXNORM, with an average of 8.6 actions (median of 2). The aggregated actions for each cluster for the two ontologies are listed in Table 6.

Comparing Browsing-Behavior Types. To visualize and investigate the differences in browsing-behavior strategies between the two ontologies, we have fit a first-order Markov chain on the extracted sequences of both projects and visualized the difference between the two transition matrices in Figure 4(b). Note that we only display the top 10 most frequent actions for reasons of readability.

CPT—being the more popular ontology—received roughly three times more requests than RXNORM, as shown in the histogram on top of Figure 4(b). The most common actions for both projects are *Browse Ontology Class* and *Browse Search.* In contrast to RXNORM, users of CPT also rely on the hierarchy of the ontology to browse classes (see *Browse Ontology Class Tree* in Figure 4(b)). This is not surprising, as RXNORM is a flat ontology, which does not have a hierarchy that can be visualized by BioPortal. This also explains why the relative importance of the transition from *Browse Search* to *Browse Ontology Class* is more dominant for RXNORM, than it is for CPT (evident in the dark brown transition probabilities in Figure 4(b)).

5. DISCUSSION

In this paper we have analyzed, modeled and clustered the browsing behavior of users on BioPortal. In particular, we have grouped users according to their browsing behavior in Section 4.1 and compared how users explore ontologies in Section 4.2.

Differences in Browsing-Behavior Types. We have shown that we were able to obtain multiple, clearly distinguishable browsing-behavior types, when using stationary distribution as input for K-means. Further, we have calculated the explained variance per cluster to determine the number of browsing-behavior types to investigate. The manual inspection of the obtained clusters shows that users primarily concentrate on browsing specific classes (*Class Explorers* and *Specific Class Browsers*) of biomedical ontologies on BioPortal. Two strategies of users to explore ontologies involve the exploitation of the ontological hierarchy (*Ontology Tree Explorers*; if available) and the search functionality provided by BioPortal (*Search Explorers*). Users in the third-largest cluster (19, 741)—BioPortal Experts—use advanced functionality, such as the annotator or the recommender, to find classes and explore ontologies.

The comparison of our results with clusters obtained from static page views indicate that the inclusion of information about the dynamic nature of a user's interaction behavior with a website is important. We were able to distill a total of 7 information consumption strategies, which we would have otherwise not identified. Note that the presented results reflect how users interact with BioPortal from a purely data-driven analysis and further analyses are required to determine the generality of our findings.

In general, our results indicate that users exhibit a preference to explore and browse the content of the classes of biomedical ontologies. However, this also indicates that additional efforts are warranted to promote and further enhance and evaluate the utility of the functionality provided by BioPortal to explore and find ontologies. For example, only a small portion of users (the BioPortal Experts) traverse along and use mappings between ontologies when visiting BioPortal. Further, the sooner we can assign a user type to any given visitor, the better we can leverage our results and automatically adapt caching and recommendation strategies to anticipate the next click.

Differences when Exploring Ontologies. Our results indicate that several characteristics of ontologies, such as the hierarchical structure (or lack thereof), can influence how users explore and interact with these semantic vocabularies and terminologies on BioPortal. One particular example of such an ontology is RXNORM, which does not exhibit a structured hierarchy, and is thus not displayed in the class explorer on BioPortal. The only ways to interact with classes of this ontology on the website is to either use the search functionality, which requires prior knowledge about the content of the ontology, or to rely on mappings from other ontologies. However, this also explains the strong focus of users/visitors of RXNORM to use the search functionality (*Browse Search*) to explore, investigate and find classes of this ontology on BioPortal (see Figure 4(b)).

In contrast, the structural hierarchy of CPT is used by BioPortal to visualize the class tree. Hence, to explore the ontology, users can (and do)—aside from searching (*Browse Search*) for specific classes—use and explore the hierarchy (*Browse Ontology Class Tree*) to find a specific class. Note that we have only presented the differences in browsing-behavior types for two of more than 500 ontologies! Yet, we have shown that simple differences (i.e., the lack of a class hierarchy) between projects cause users of BioPortal to employ different strategies to find classes or explore the ontological content. We leave it open for future work to continue this line of investigation to identify characteristics and features of ontologies causing differences in how users seek and consume information stored inside them.

6. CONCLUSIONS & FUTURE WORK

In this paper, we have presented an approach for identifying different types of user behavior in ontology exploration log data using stationary distributions from first order Markov chains. We provide evidence for a total of 7 distinct browsing-behavior types on BioPortal and offer an interpretation of their relevance and meaning. In addition, we cluster ontologies by their user behavior and identify pragmatic commonalities among ontology projects. Our results advance our understanding of the ways in which ontological repositories, such as BioPortal, are used and may guide the development of more user-oriented systems for ontology exploration and reuse on the Web.

For future work, we are particularly interested in further elaborating the comparison between the browsing behavior of different ontologies on BioPortal. In doing that, we might be able to identify commonalities for groups of ontologies that trigger differences in the way users interact with ontologies on BioPortal. Understanding the different ways in which semantic structures may influence ontology exploration behavior has the potential to help in the design of BioPortal as well as in the development of more effective and more reusable ontologies. Additionally, we would like to compare the results presented in this paper to usability studies, testing the utility of different features in BioPortal and how they—on a qualitative level—influence the browsing behavior of users.

7. ACKNOWLEDGMENTS

This work was supported by Grant GM086587 from the U.S. National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health. The Protégé project is supported by NIGMS grant GM103316. NCBO is funded by National Institutes of Health grant U54 HG004028.

8. REFERENCES

[1] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[2] J. Borges and M. Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Trans. on Knowl. and Data Eng.*, 19(4):441–452, Apr. 2007.

[3] F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlos. Are web users really markovian? In *Proceedings of the 21st international conference on World Wide Web*, pages 609–618. ACM, 2012.

[4] M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. *ACM Trans. Internet Technol.*, 4(2):163–184, May 2004.

[5] L. Espín Noboa, F. Lemmerich, P. Singer, and M. Strohmaier. Discovering and characterizing mobility patterns in urban spaces: A study of manhattan taxi data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 537–542, 2016.

[6] S. M. Falconer, T. Tudorache, and N. F. Noy. An analysis of collaborative patterns in large-scale ontology development projects. In M. A. Musen and O. Corcho, editors, *K-CAP*, pages 25–32. ACM, 2011.

[7] B. Fortuna, D. Mladenic, and M. Grobelnik. User modeling combining access logs, page content and semantics. *arXiv preprint arXiv:1103.5002*, 2011.

[8] A. Gomez, M. Fernandez, and O. Corcho. Ontological engineering. *2 nd Edition, Springer-Verlag*, 2004.

[9] Z. Hellman and A. Gal. *Structural Conflict Avoidance in Collaborative Ontology Engineering*, pages 216–225. Springer Netherlands, Dordrecht, 2005.

[10] J. Hoxha, M. Junghans, and S. Agarwal. Enabling semantic analysis of user browsing patterns in the web of data. *arXiv preprint arXiv:1204.2713*, 2012.

[11] S. Jupp, T. Burdett, J. Malone, C. Leroy, M. Pearce, J. McMurry, and H. Parkinson. A new ontology lookup service at embl-ebi.

[12] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, pages 561–570. ACM, 2010.

[13] C. Lakshminarayan, R. Kosuru, and M. Hsu. Modeling complex clickstream data by stochastic models: Theory and methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 879–884, 2016.

[14] J. Lehmann, C. Müller-Birn, D. Laniado, M. Lalmas, and A. Kaltenbrunner. Reader preferences and behavior on wikipedia. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 88–97. ACM, 2014.

[15] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Comput. Netw.*, 33(1-6):387–401, June 2000.

[16] M. A. Musen, N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M.-A. Story, B. Smith, et al. The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, 19(2):190–195, 2012.

[17] N. F. Noy, D. L. McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[18] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440, 2009.

[19] C. Pesquita and F. M. Couto. Predicting the extension of biomedical ontologies. *PLoS Comput Biol*, 8(9):e1002630, 09 2012.

[20] P. L. T. Pirolli and J. E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, 1999.

[21] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy. Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic web*, 4(3):277–284, 2013.

[22] R. Sen and M. Hansen. Predicting a web user's next access based on log data. *Journal of Computational Graphics and Statistics*, 12(1):143–155, 2003.

[23] E. Simperl and M. Luczak-Rösch. Collaborative ontology engineering: a survey. *The Knowledge Engineering Review*, 29(01):101–131, 2014.

[24] P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1003–1013, 2015.

[25] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[26] M. Strohmaier, S. Walk, J. Pöschko, D. Lamprecht, T. Tudorache, C. Nyulas, M. A. Musen, and 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), 2013.

[27] S. Van Laere, R. Buyl, and 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*, pages 657–673. Springer, 2014.

[28] S. Walk, P. Singer, L. E. Noboa, T. Tudorache, M. A. Musen, and 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*, pages 551–568, 2015.

[29] S. Walk, P. Singer, and M. Strohmaier. Sequential action patterns in collaborative ontology-engineering projects: A case-study in the biomedical domain. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1349–1358, 2014.

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

[31] S. Walk, P. Singer, M. Strohmaier, T. Tudorache, M. A. Musen, and N. F. Noy. Discovering beaten paths in collaborative ontology-engineering projects using markov chains. *Journal of Biomedical Informatics*, 51:254–271, 2014.

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

[33] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. 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.