The Provenance Bee Wiki: Tracking the Growth of Semantic Wiki Communities

Yolanda Gil, Dipsy Kapoor, Reed Markham, and Varun Ratnakar
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey CA, 90292
gil@isi.edu, dipsy@isi.edu, rmarkham@usc.edu, varunr@isi.edu

ABSTRACT
Contributors in hundreds of semantic wiki sites are creating structured information in RDF every day, thus growing the semantic content of the Web in spades. Although wikis have been analyzed extensively, there has been little analysis of the use of semantic wikis. The Provenance Bee Wiki was created to gather and aggregate data from these sites, show how this content is growing over time, and to make all this detailed data readily available to the research community. We also present a high-level analysis of the almost 600 wikis indexed in Provenance Bee Wiki that have less than 5,000 pages.

Author Keywords
Semantic wikis; semantic web; RDF; social knowledge collection.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces - Interaction styles.

INTRODUCTION

Semantic wikis allow a community of users to structure information by creating concepts and properties that link across wiki topic pages [Bry et al 12; Krötzsch et al 2007]. Semantic wikis are a very successful platform for social knowledge collection [Gil 2014]. There are hundreds of communities that are using semantic wikis for social knowledge collection. They are collaboratively creating structured content. Some semantic wikis have a serious use, such as scientific knowledge organization. Others have practical use, for example gardening or restaurant finding. Some wikis have users that are on the younger side, such as teens who organize information about the different characters in the cards they trade. Perhaps the best known semantic wiki is Wikidata [Vrandečić and Krötzsch 2014], a structured version of Wikipedia.

Although there has been a significant amount of work studying wiki communities (notably Wikipedia), there is very little work on analyzing communities that generate structured content, such as semantic wikis. Prior work analyzed semantic wiki communities in terms of the organization of the contributors and the evolution of structured content over time [Gil and Ratnakar 2013; Gil et al 2013a; Gil et al 2013b]. Using data from 230 semantic wikis about their creation and edits, as well as the wiki users who edited them, the study found that 1) concepts are not used very often, and not used at all in many wikis; 2) properties were used in all the wikis, although very small numbers of users edit them; and 3) very large numbers of property assertions are used in almost every wiki. However, the study was limited because the data that we collected was for a subset of all semantic wikis, focused on a particular time period, and focused on a particular set of questions that required data about the provenance of the contributions. It would be very useful if this kind of study was more common and repeatable.

Our goal is to make data about semantic wiki communities readily available so that it becomes very easy to do studies about their organization and the dynamic evolution of the wiki's contents.

This paper presents our work on Provenance Bee Wiki, a community resource for studying semantic wiki data. It is a central repository with detailed provenance data of edits in semantic wikis. The site includes aggregated statistics that highlight the most salient wikis as well as the total amount of semantic content across all wikis. We also present a high-level analysis using data from the Provenance Bee Wiki for 594 semantic wikis of smaller size (5,000 pages or less), focusing on how users edit the wiki to add structure to the contents. We analyze the concepts and properties created, the amount of editors involved in creating them, and the amount of edits in each category.

K-CAP 2015, October 07 - 10, 2015, Palisades, NY, USA

ACM 978-1-4503-3849-3/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2815833.

RELATED WORK

[Bry et al 12] gives a detailed overview of semantic wikis and a thorough comparison of alternative implementations, including Semantic MediaWiki, OntoWiki, and AceWiki.

While there are many published analyses of wikis (e.g., [Kittur et al 2009; Leskovec et al 2010]), semantic wiki communities have not been studied in depth. Some prior work has concerned analyzing semantic wikis communities. Some work has focused on the success or failure of communities in terms of reaching critical mass [Walk et al 2013], but it has looked at general user data in the wikis without focusing on semantic aspects of the contributions. Our prior work addressed how users edit wikis to add structure to their contents [Gil and Ratnakar 2013; Gil et al 2013a; Gil et al 2013b], as mentioned in the introduction. Although this was the first comprehensive study of semantic wiki communities, it used a limited number of semantic wikis and a specific time period. Our work with the Provenance Bee Wiki seeks to provide more comprehensive data.

Structured knowledge collection from volunteers has been studied in prior work, including OpenMind [Lieberman et al 2004], the Cyc FACTory [Matuszek et al 2005], and Learner [Chklovski and Gil 2005]. In contrast with semantic wikis, contributors to these sites are not involved in the use of content nor do they interact as a community.

Collaborative ontology editors allow communities of users to create classes and properties that describe a domain. A popular collaborative ontology editor is Collaborative Protégé [Tudorache et al 2011; Tudorache and Musen 2011]. A study of collaborative ontology editors found that the types of contributions are diverse and shape the role of the contributors [Strohmaier et al 2013]. Unlike with semantic wikis, which do not require much training, users have to be trained to use those editors.

OVERVIEW OF THE PROVENANCE BEE WIKI

The Provenance Bee Wiki collects information about the structure contents of existing semantic wikis, together with their provenance in terms of who added the content and when. The Provenance Bee Wiki is publicly available and its data accessible to the community [PBW 2015]. The Provenance Bee Wiki focuses on data from Semantic MediaWikis [Krötzsch et al 2007], since their contents are easily accessible through APIs[1].

Architecture

Figure 1 shows an overview of the architecture of the Provenance Bee Wiki. The rest of this section describes each aspect of the system in detail.

The Provenance Bee Wiki gets basic information from WikiApiary[2], a site that already exists and is a repository of

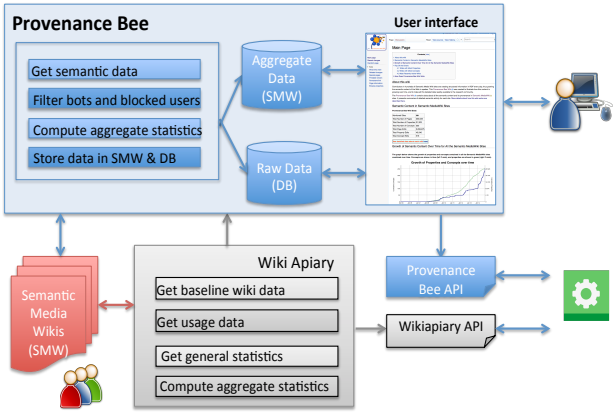

Figure 1. Overview of the architecture of the Provenance Bee Wiki.

data about WikiMedia sites. The Provenance Bee Wiki also accesses semantic wikis directly to collect more detailed data together with its provenance. It then computes aggregate statistics based on this detailed data.

The Provenance Bee Wiki itself is implemented as an extension of Semantic MediaWiki. The detailed data is stored in a database, while the aggregate statistics for each wiki are stored as semantic properties of that wiki.

The Provenance Bee Wiki, as any wiki, has a user interface that allows users to browse the data. Total statistics for all wikis are shown in graphs. For each wiki, detailed information about the growth of its structured content and its user base are shown.

The data collected by the Provenance Bee Wiki is available programmatically through an API, as is the data from WikiApiary.

Availability of Data and Software

All the data analyzed in this paper can be found at [PBW 2015a]. The software is also available at [PBW 2015b].

Accessing Wiki Data

The data of the semantic wikis are accessible through APIs. All MediaWikis have an API that allows accessing data about users, content, and edits together with time stamps. Semantic MediaWikis have an extended API to access to edits of semantic properties and concepts.

The Provenance Bee Bot

The Provenance Bee is an extension of WikiApiary, a centralized repository of summary information for MediaWiki (and therefore Semantic MediaWiki) sites. WikiApiary itself is a Semantic MediaWiki site.

WikiApiary has a list of wiki sites that it analyzes. For each MediaWiki site, a page of type "Website" is created. For each Website some semantic properties are included,

[1] https://semantic-mediawiki.org/
[2] https://wikiapiary.com/

such as API URL, site URL, isDefunct (not accessible for some period of time), and isInactive.

WikiApiary regularly runs a bot, called Bumble Bee, which reads the list of registered wikis and filters out wikis that are inactive or not accessible. Then for each wiki site, WikiApiary makes calls to the wiki APIs to collect general site information, site statistics like the number of pages page edits, and page editors. It also collects some basic Semantic MediaWiki statistics. All this information is represented using the semantic properties mentioned above.

The Provenance Bee is the bot that populates the content of the Provenance Bee Wiki. The Provenance Bee extends WikiApiary to collect timestamp and editor information for all revisions for all pages, concepts and properties.

Given the list of semantic wikis that it gets from WikiApiary, the Provenance Bee accesses each wiki and gets the following aggregate information through the API:

- Number of properties
- Number of declared properties
- Number of used properties
- Number of pages
- Number of concepts
- Number of assertions
- Number of days between first and last wiki edits
- Number of days since last wiki edit

These are represented as semantic properties for that wiki.

For those wikis that allow detailed queries about their content, the Provenance Bee gathers detailed data about:

- The semantic content of the wiki: each concept, property, and assertion at each point in time
- The contributions of each user: the content and edits made by each user at each point in time

These data are stored in a SQL database.

The Provenance Bee aggregates these detailed data to generate the following statistics about the wiki:

- Number of page editors
- Number of property editors
- Number of concept editors
- Number of page edits
- Number of property edits
- Number of concept edits

These are represented as semantic properties for each wiki.

The Provenance Bee aggregates the data and stores the following metrics:

- Daily number of page edits
- Daily number of property edits
- Daily number of concept edits

The Provenance Bee also calculates normalized counts to make the wikis more comparable:

- Proportion of the total page edits that were made by a given editor
- Proportion of the total property edits that were made by a given editor

The Provenance Bee also collects bot and blocked user information in order to filter out spam data. Many wikis use bots to populate their contents, typically making assertions based on already defined properties. This helps the Provenance Bee keep track of human editors and extract more accurate information about their edits.

The Provenance Bee extracts data continuously, and it saves historical data so data only has to be extracted once.

Programmatic Access to Provenance Bee Data

Provenance Bee stores all the data it extracts as semantic content in the Provenance Bee Wiki. The provenance summary for each wiki is asserted as semantic properties of wiki sites that can be queried using the Semantic MediaWiki API.

All the data in the Provenance Bee Wiki is stored in RDF and therefore the data can be queried.

Wiki content can be generated dynamically from the data through queries that generate tables. These tables can be displayed within wiki pages when the query is embedded in the wiki page. For example, the following Semantic MediaWiki ask query will return a table containing the page, property and concept counts for all wikis:

```
[[Category:Website]][[Has provenance
info::True]]|?Has ID|?Has prov page
count|?Has prov property count|?Has
prov concept count
```

The data can also be accessed programmatically via a SPARQL API.

Browsing Provenance Bee Data

The Provenance Bee Wiki interface shows aggregate data across all the wikis, such as the total amount of properties and concepts as well as total amount of edits.

The Provenance Bee Wiki interface also shows data for each of the wikis, and includes some of the WikiApiary information.

Coverage of Provenance Bee Data

Provenance Bee accesses the list of wikis from WikiApiary and selects those that are Semantic Media Wiki sites in order to make its own list of wikis.

Some wikis are very large, in the order or millions of pages. Accessing the detailed history data for all the pages one by one via the API is not very practical. The queries have to be managed and tracked, as wiki connections often time out and a given query has to be reissued if interrupted. In addition, the wikis sometimes disable access when hit by so many queries, otherwise their performance could be affected. Therefore, the Provenance Bee Wiki accesses only

Table 1. Overall statistics of the semantic wikis tracked by the Provenance Bee Wiki.

	June 2015	June 2014
Monitored Sites	594	591
Total Number of Pages	727,110	353,535
Total Number of Properties	22,467	21,323
Total Number of Concepts	433	235
Total Page Edits	2,771,395	2,293,975
Total Property Edits	45,501	43,540
Total Concept Edits	820	516

a subset of wikis of manageable size. It currently selects wikis with less than 5,000 pages, for a total of 594 wikis. We are designing a more scalable approach to extract data from larger wikis.

ANALYZING PROVENANCE BEE DATA

This section gives an overview of the data currently available as of the writing of this paper. The Provenance Bee bot continues to gather additional data about new wikis as they become available as well as increments the data about the wikis it already tracks.

As mentioned earlier, the Provenance Bee Wiki tracks wikis that have less than 5,000 pages. At the time of writing this article, there are 723 such wikis. However, many of them require a login to access the API, so the Provenance Bee Wiki does not collect their data. The Provenance Bee Wiki currently tracks a total of 594 wikis.

Table 1 shows summary statistics for all the semantic wikis in 2015 and 2014. This table is dynamically generated and available in the front page of the Provenance Bee Wiki. We note that the content continues to grow, particularly the amount of semantic properties and concepts.

Note that the amount of properties created is over 22,000. This is quite a large number, and it shows that this feature of semantic wikis is heavily used. On the other hand, there is very limited use of concepts in the wikis, since only 433 concepts are defined in total. These findings are consistent with the findings of our earlier work [Gil and Ratnakar 2013; Gil et al 2013a; Gil et al 2013b], and they now apply to a larger number of wikis (although our original study included less wikis it included wikis of larger sizes). However, there is an 84% increase in the amount of concepts created in just one year in the same amount of wikis, so the creation of concepts appears to be on the rise.

Figure 2 shows the data collected for a particular semantic wiki, the Proteomics Wiki. All these graphs, as well as summary statistics about the wiki, and a pointer to WikiApiary (which in turn offers pointers to the wiki itself as well as its API) are offered for every wiki in its own page of the Provenance Bee Wiki.

Figure 3 shows aggregate data for all the semantic wikis tracked by the Provenance Bee Wiki. The top graph shows the growth of properties and concepts contained in all the Semantic MediaWiki sites combined over time. Concepts are shown in blue (left Y-axis), and properties are shown in green (right Y-axis). The bottom graph shows the number of property edits and concept edits over time in all the of all the Semantic MediaWiki sites combined. Concepts are shown in blue (left Y-axis), and properties are shown in green (right Y-axis).

The rest of this section presents a high-level analysis of the data for the 594 wikis in the Provenance Bee Wiki. Note that the analysis focuses on wikis that have less than 5,000 pages since those are the wikis currently targeted by the Provenance Bee Wiki, while prior work [Gil et al 2013] analyzed 230 wikis but they were of all sizes. Therefore, our data is not comparable because it is about different (though intersecting) sets of wikis.

Concept Editors

The vast majority of wikis in our sample (~92%) had no concepts defined. 13 wikis had numbers of concepts ranging between 5 and 99, with a mean of about 30. Only 9 wikis in this sample have multiple concept editors, 39 wikis have only 1 editor, and the remaining 546 have none.

Concept Edits

Only two wikis had over 100 edits, and 17 wikis had between 10 and 99 concept edits, compared to 19 wikis in the last analysis.

Property Definitions

We found that only 18 total wikis have more than 200 properties defined. Only 28% of wikis have no semantic properties defined, in contrast with the high number of wikis that have no concepts defined.

We also looked at the proportion of semantic properties defined in the wiki over the total amount of pages. The range of values varies between 0.01 and 10, with a mean of 0.32 and standard deviation of 0.71. Most wikis have value of less than 1, which supports the notion that some pages have no properties defined. Several wikis have a value greater than 1, which indicates that many pages include several properties. There is a significant outlier with a value of 136.

Property Editors

Another important aspect of the semantic uses of the wikis is the number of contributors that create and change semantic properties. Only 11 wikis have more than 10 property creators. There are 3 wikis in our sample with at least two-dozen distinct property editors.

Property Edits

Another important aspect of the semantic uses of the wikis is the number of semantic property edits. 20 of the wikis have more than 400 property edits. There are 154 wikis that have no ongoing property edits on their pages.

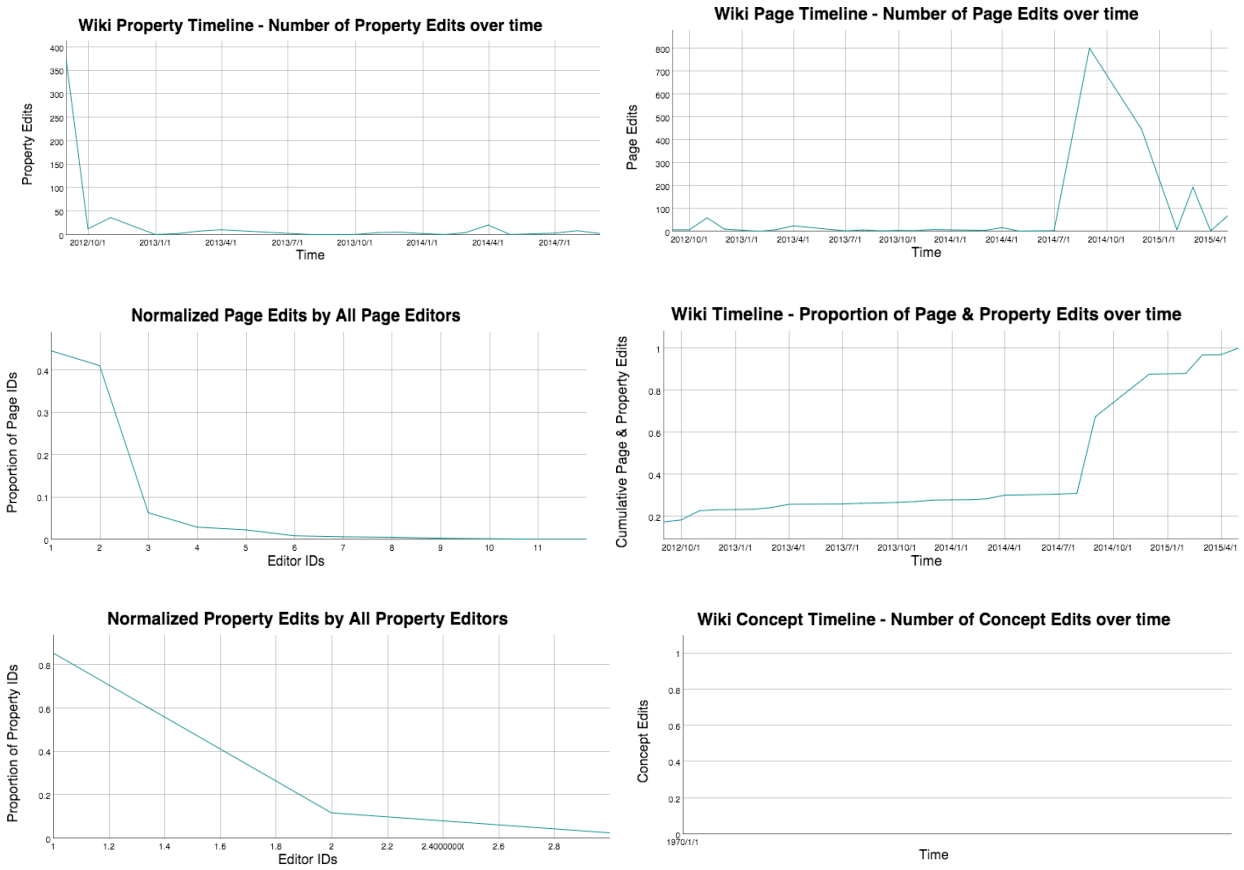

Figure 2. Data for a particular Semantic MediaWiki (the Proteomics Wiki) shown in the Provenance Bee Wiki.

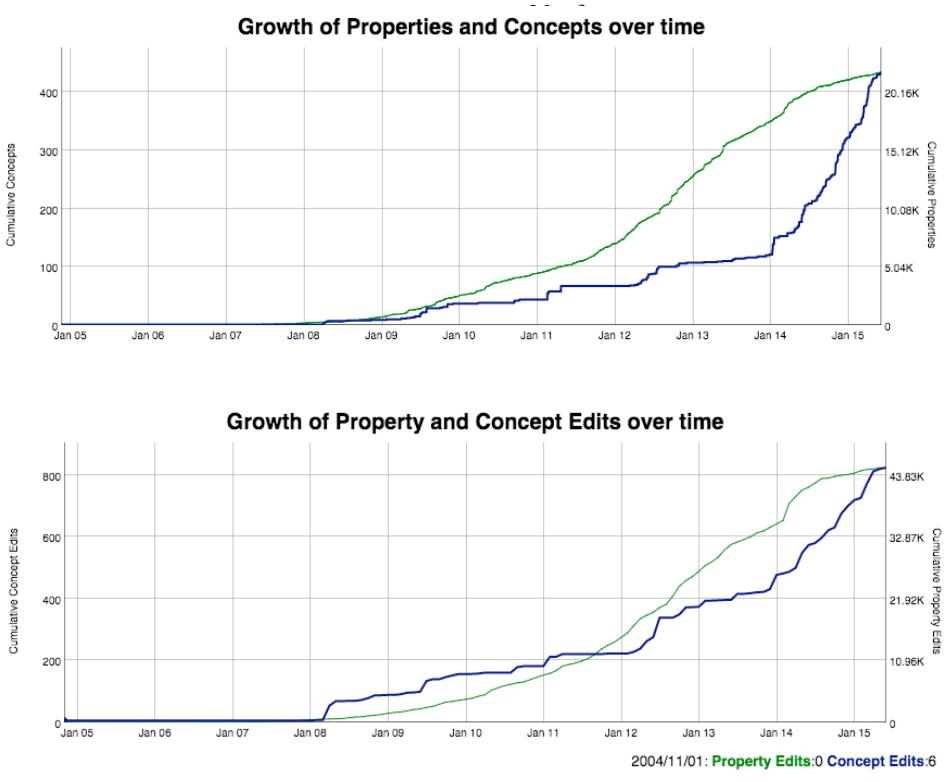

Figure 3. Aggregate statistics about all Semantic MediaWikis in the Provenance Bee Wiki.

The lifespans of the semantic wikis in our sample are highly variable, ranging from 24 days to over 14 years. A semantic wiki in our sample has an expected lifespan of over 4 years. On average, the wikis that have been dormant is for about 232 days, with a maximum of almost 7 years.

Figures 4 and 5 show lifespan data, in logarithmic scales. Figure 4 shows number of pages, number of page editors, and number of page edits against lifespan. They are generally correlated with increases in lifespan. Figure 5 shows number of properties, number of property editors, and number of property edits versus lifespan. They are generally correlated with increases in lifespan.

Figure 6 and 7 show page editing activity and property editing activity respectively for comparison purposes, both in logarithmic scales. Figure 6 shows the page editors against total pages and page edits, and the total pages against the page edits. Figure 7 shows corresponding data for semantic properties: property editors against total properties and property edits, and the total properties against the property edits. While a greater amount of total pages and total edits seem correlated with a longer lifespan of wikis, the same trend is not the case for semantic properties. Wikis with smaller amounts of properties, editors, and property edits have comparable lifespan to larger wikis. We also note the amount of property editors in a wiki does not correlate with lifespan.

Figure 6 shows that the amount of page editors does not correlate with the amount of pages or page edits. Figure 7 shows a similar situation for property editors, although there are much fewer numbers of property editors than page editors. Semantic wikis tend to have few property editors (consistent with the [Gil et al 2013 study]), but they make a widely ranging amount of edits in different wikis. This means that a large amount of semantic content can be created by a few editors. For both wiki pages and semantic properties, it is apparent that growth in content (total pages and total properties) is strongly correlated with growth in editing activity (page edits and property editors).

DISCUSSION

Although we only analyzed semantic wikis that were of smaller size, some interesting trends can be highlighted from the data reported in this paper:

- *Semantic wiki communities continue to grow the semantic content of the Web.* The amount of new concepts and properties being created and the amount of edits to existing ones continues to grow.
- *Semantic wiki communities are creating more concepts.* While there has been an increase in the number of properties created, the number of concepts created has almost doubled.
- *Most wikis have a small number of property and concept editors*. The majority of semantic wikis have few people editing the semantic structures of the wikis.
- *The lifespan of wikis does not appear to be correlated with the amount of semantic content*. We see no evidence that communities that adopt semantic wikis have short lifespans.
- *Contributors that create semantic properties can add large amounts of content*. Although most semantic wikis have very few property editors, they can be responsible for a large amount of edits to properties.
- *Semantic content appears to evolve just as the rest of the content in wikis.* There is a strong correlation with the amount of properties and the amount of property edits. There is also a strong correlation between amount of pages and amount of page edits as well, so properties appear to evolve as the rest of the content.

CONCLUSIONS AND FUTURE WORK

The Provenance Bee Wiki was designed to track the provenance of the semantic content of Semantic MediaWiki sites. At the moment, it tracks provenance in terms of what users contribute which content over time. We collected content for almost 600 wikis, all with 5,000 pages of less. We report on a continuous growth of semantic content in these wikis since we started tracking them in 2014. We also detected an increased use of concepts, significant activity by property editors even where there are few, and a co-evolution of the edits to regular pages and to properties.

In the future, we plan to extract more fine-grained provenance details from the original Semantic MediaWiki sites. Some of the provenance information is hard to access, and may require changes in how the Semantic MediaWiki platform tracks updates.

We are currently analyzing these behaviors in more detail. We are trying to understand the reasons for the small number of users who make property definitions. This may be a result of use and enforcement of restrictive editing policies. It is possible that additional users would be involved in the creation of properties if there were facilities in the wiki to detect and resolve conflicts collaboratively.

We also plan to carry out analyses that focus on the evolution of the wikis over time. This would lead to a better understanding of how wiki communities form and evolve, and how the characteristics of the community and its behaviors influence their lifespan and evolution.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from the US National Science Foundation with grant IIS-1117281.

REFERENCES

[Bry et al 2012] François Bry, Sebastian Schaffert, Denny Vrandečić and Klara Weiand. "Semantic Wikis: Approaches, Applications, and Perspectives." Lecture Notes in Computer Science, Reasoning Web. Semantic Technologies for Advanced Query Answering, Vol. 7487, 2012.

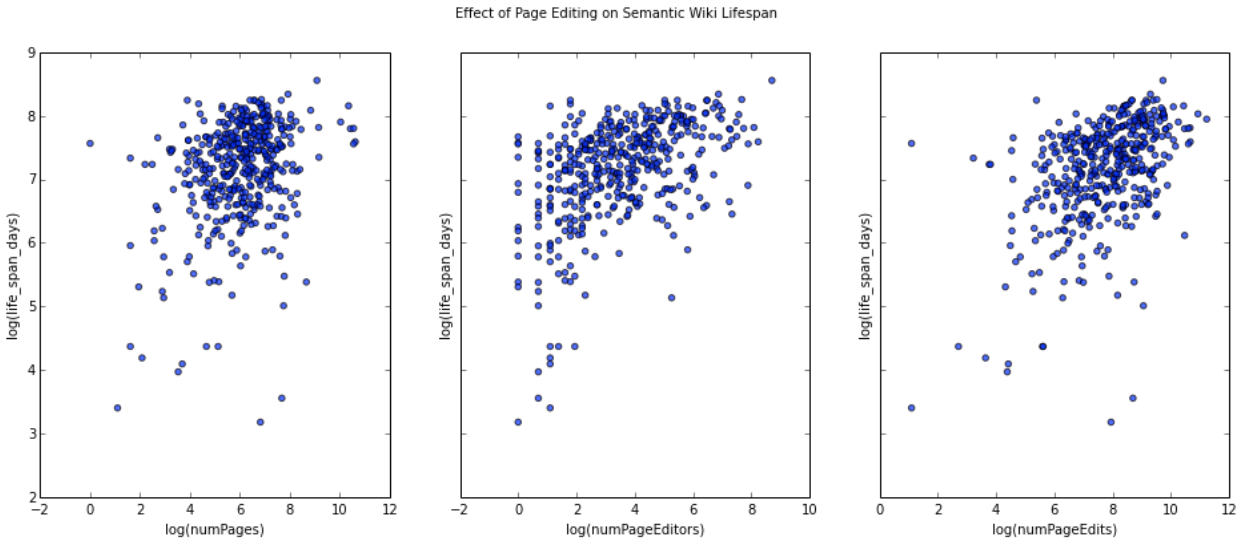

Figure 4. Pages, page editors, and page edits versus life span of wikis

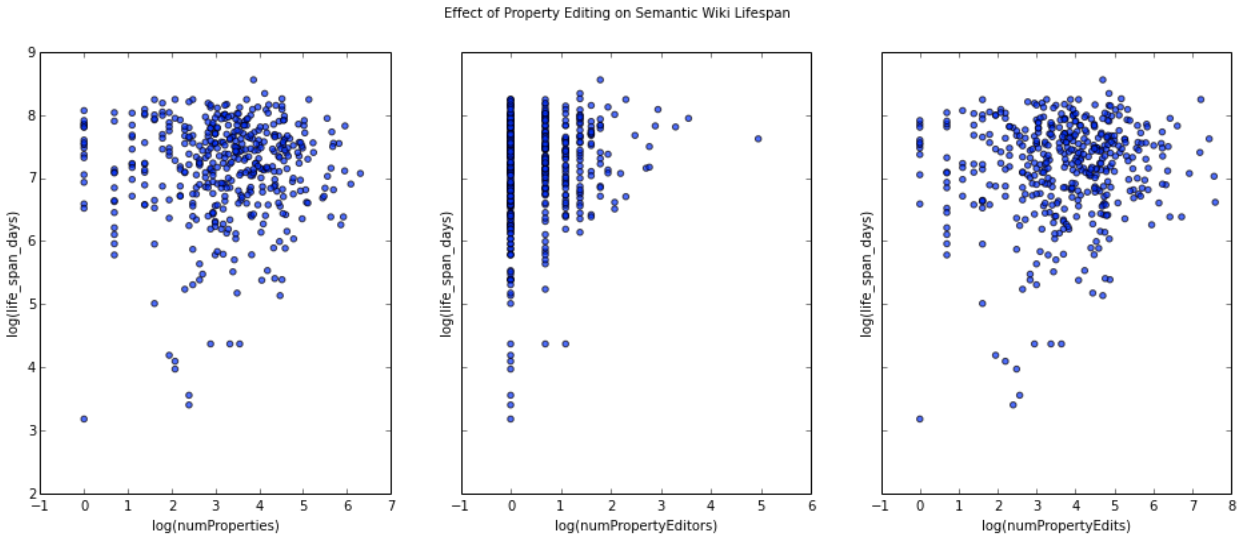

Figure 5. Properties, property editors, and property edits versus life span of wikis.

[Chklovski and Gil 2005] Tim Chklovski and Yolanda Gil. "An Analysis of Knowledge Collected from Volunteer Contributors." Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI), 2005.

[Gil 2014] Yolanda Gil. "Social Knowledge Collection." In Handbook of Human Computation. P. Michelucci (Ed). Springer, 2013.

[Gil and Ratnakar 2013] Gil, Y.; and Ratnakar, V. "Knowledge Capture in the Wild: A Perspective from Semantic Wiki Communities." Proceedings of the Seventh ACM International Conference on Knowledge Capture (K-CAP), 2013.

[Gil et al 2013a] Yolanda Gil, Angela Knight, Kevin Zhang, Larry Zhang, and Ricky Sethi. "An Initial Analysis of Semantic Wikis." Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI), 2013.

[Gil et al 2013b] Gil, Y.; Knight, A.; Zhang, K.; Zhang, L.; Ratnakar, V.; and Sethi, R. "The Democratization of Semantic Properties: An Analysis of Semantic Wikis." Proceedings of the Seventh IEEE International Conference on Semantic Computing (ICSC), 2013.

[Kittur et al 2009] Aniket Kittur, Bryant Lee, Robert E. Kraut. "Coordination in collective intelligence: the role of team structure and task interdependence." Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI), 2009.

[Krötzsch et al 2007] Markus Krötzsch, Denny Vrandečić, Max Völkel, Heiko Haller, Rudi Studer. "Semantic Wikipedia." Journal of Web Semantics, 5(4), 2007.

[Leskovec et al 2010] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. "Governance in Social Media: A case study of the Wikipedia promotion process." Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM), 2010

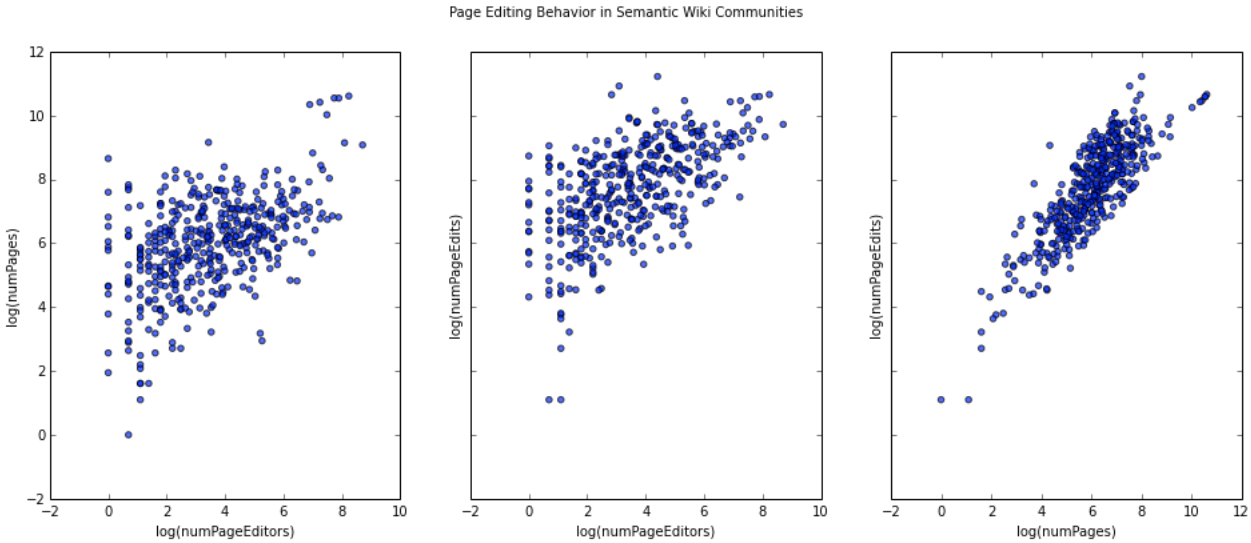

Figure 6. Page editors against total pages and page edits, and total pages against page edits.

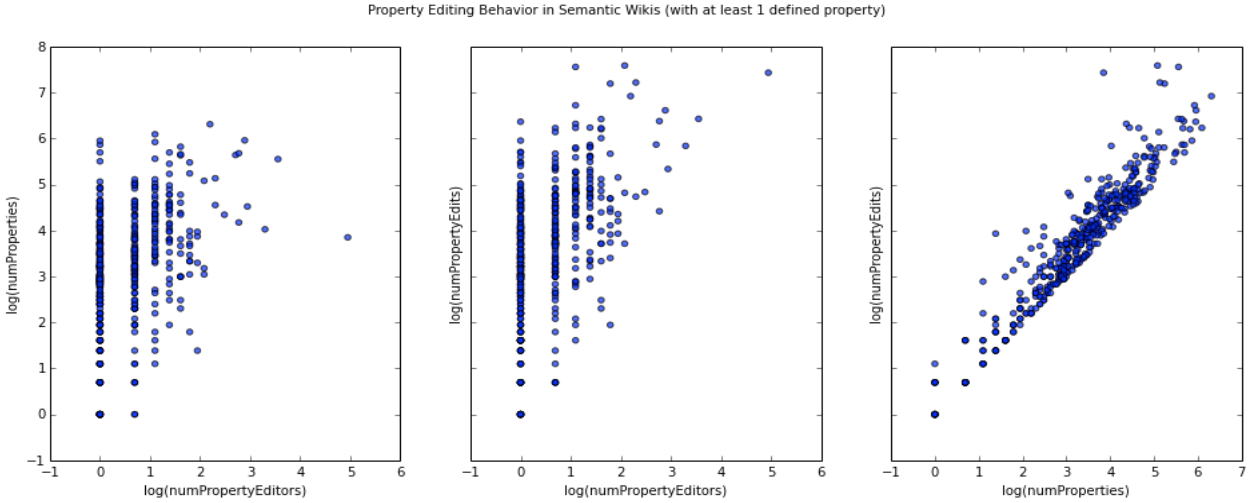

Figure 7. Property editors against total properties and property edits, and total properties against property edits.

[Lieberman et al 2004] Henry Lieberman, Hugo Liu, Push Singh, and Barbara Barry. "Beating some common sense into interactive applications." AI Magazine, 25(4), 2004.

[Matuszek et al 05] C. Matuszek, M. J. Witbrock, R. C. Kahlert, J. Cabral, D. Schneider, P. Shah, D. B. Lenat. "Searching for Common Sense: Populating Cyc from the Web". Proceedings of the 20th National Conference on Artificial Intelligence (AAAI), 2005.

[PBW 2015a] Provenance Bee Wiki Data. Available from http://skc.isi.edu/provenancebeewiki/. Last accessed 20 August 2015.

[PBW 2015b] Provenance Bee Wiki Software. Available from https://github.com/IKCAP/ProvenanceBee. Last accessed 20 August 2015.

[Strohmaier et al 2013] M. Strohmaier, S. Walk, J. Poeschko, D. Lamprecht, T. Tudorache, C. Nyulas, M. Musen, and N.F. Noy. "How ontologies are made: Evaluation of the hidden social dynamics behind collaborative ontology engineering projects." Journal of Web Semantics, 20, 2013.

[Tudorache and Musen 2011] T. Tudorache, M. A. Musen. "Collaborative Development of Large-Scale Biomedical Ontologies." In Collaborative Computational Technologies for Biomedical Research, S. Ekins, M. A. Z. Hupcey and A. J. Williams, John Wiley & Sons, 2011.

[Tudorache et al 2011] T. Tudorache, C. I. Nyulas, M. A. Musen, N. F. Noy. "WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web." Semantic Web Journal, 11-165, 2011.

[Vrandečić and Krötzsch 2014] Denny Vrandečić, Markus Krötzsch. "Wikidata: A Free Collaborative Knowledgebase." Communications of the ACM 57:10, 2014.

[Walk et al 2013] S. Walk, J. Pöschko, M. Strohmaier, K. Andrews, T. Tudorache, N. F. Noy, C. I. Nyulas, M. A. Musen. "PragmatiX: An Interactive Tool for Visualizing the Creation Process Behind Collaboratively Engineered Ontologies." International Journal on Semantic Web and Information Systems, 2013.