Harvesting Wiki Consensus

Using Wikipedia Entries as Vocabulary for Knowledge Management

Vocabularies that provide unique identifiers for conceptual elements of a domain can improve precision and recall in knowledge-management applications. Although creating and maintaining such vocabularies is generally hard, wiki users easily manage to develop comprehensive, informal definitions of terms, each one identified by a URI. Here, the authors show that the URIs of Wikipedia entries are reliable identifiers for conceptual entities. They also demonstrate how Wikipedia entries can be used for annotating Web resources and knowledge assets and give precise estimates of the amount of Wikipedia URIs in terms of the popular Proton ontology's top-level concepts.

Martin Hepp, Katharina Siorpaes, and Daniel Bachlechner
Digital Enterprise Research Institute, University of Innsbruck

Knowledge management aims to help organizations and individuals better exploit their intellectual assets — particularly by reusing previous experiences and improving access to knowledge distributed over multiple human actors, systems, and other resources. Retrieving relevant assets can be difficult because the conceptual specificity of terms in a search task is frequently very high. Also, an organization's most valuable assets often occupy areas with high conceptual dynamics due to innovation, which means it must be possible to add novel elements to the vocabulary in a timely manner.

Ontologies — consensual, explicit conceptualizations of a domain of discourse[1–3] — are a candidate technology for improving precision and recall in knowledge management. Unfortunately, potential adopters of ontology-based solutions face a severe shortage of current, high-quality ontologies for many domains. Many ontologies published on the Web are outdated, "dead" collections that single individuals created in some academic research context. One potential explanation for this is that creating and maintaining an ontology requires specific tools and skills, which domain experts frequently lack. In contrast, wikis make it very simple for individuals to create new entries or to modify existing ones, and

1089-7801/07/$25.00 © 2007 IEEE

they're extremely popular. For instance, the English version of Wikipedia (http://en.wikipedia.org/) contains more than 1.5 million entries (as of 13 January 2007) contributed by a very broad range of Web users, and is growing at an amazing pace. This means that Wikipedia also provides URIs as unique identifiers for the same amount of topics.

Here, we regard wikis' infrastructure and culture as an environment for constructing and maintaining consensual vocabularies and suggest using the Wikipedia URIs as identifiers for conceptual entities for annotating knowledge assets. Through a quantitative analysis, we show that Wikipedia URIs are surprisingly reliable for this purpose. We've also estimated the proportion of Wikipedia URIs in terms of the Proton ontology's top-level concepts (see http://proton.semanticweb.org), which reveals the type and ontological nature of the available conceptual elements.

Ontologies vs. Vocabularies

Ontologies are unambiguous representations of concepts, relationships between concepts (such as a hierarchy), ontologically significant individuals, and axioms. (A comprehensive overview of ontologies is available elsewhere.[3]) "Unambiguous" in this context means that ontologies let users grasp any element's meaning so that they understand the vocabulary when annotating data or expressing queries. Also, they have a formal semantics to support machine reasoning and to explicitly exclude unwanted interpretations (from a logician's perspective, this is the most important property).

However, ontologies aren't just formal representations of a domain — they're *community contracts* about such representations. Given that a discourse is a dynamic, social process during which participants often modify or discard previous propositions or introduce new topics, such a community contract can't be static, but must evolve. Also, the respective community must be technically and skill-wise able to build or commit to the ontology.[4]

In computer science, we usually assume that we can define ontologies' conceptual entities mainly by formal means — for example, we use axioms to specify the intended meaning of domain elements. We usually consider conceptualizations that provide domain elements defined using only informal means (such as natural language) to be only vocabularies, not ontologies. In contrast, in information systems, researchers discussing ontologies are more concerned with understanding conceptual elements and their relationships, and often specify their ontologies using only informal means, such as UML class diagrams, entity-relationship models, semantic nets, or even natural language. In such contexts, a collection of named conceptual entities with a natural language definition — that is, a controlled vocabulary — would count as an ontology. To avoid confusion, we'll use the term "vocabulary" in this article rather than "lightweight ontology."

In general, having a few skilled individuals carefully construct the representation of a domain of discourse for a larger user community is problematic. Maintenance isn't under that community's control,[4] so users can't add missing entries if they spot the need for a new concept and must rely on a small group of privileged creators. Thus, users might not report missing entries at all, and additions might take too long in quickly evolving domains. In contrast, with natural language, the vocabulary's evolution is under the user community's control — anybody can invent and define a new word or concept in the course of communication.

An indicator for this problem is the gap in popularity between Web 2.0 techniques — namely, tagging, folksonomies,[5] and vocabularies such as friend-of-a-friend (FOAF) — and engineered, formal ontologies. We assume that the ease with which users can be involved in wikis, combined with the use of URIs as identifiers for wiki pages, makes wikis a promising platform for developing vocabularies for knowledge management.

Using Wikipedia Entries as a Vocabulary for Annotations

We suggest using Wikipedia, and wiki implementations in general, as a means for

- defining URIs for conceptual entities,
- describing those entities' meanings in natural language, augmented by multimedia elements (drawings, pictures, videos, or sound recordings), and
- preserving the discourse that led to an entry's current version as an important part of the respective conceptual entity's definition.

A standard wiki already provides all the functionality necessary for creating a textual definition and a unique URI. We could immediately use this mechanism and reuse the URI not only as a resource locator for retrieving the description but also as the identifier for the respective conceptual entity. Our approach was motivated by the following considerations:

- Wikipedia is the biggest available collection of conceptual elements defined by a textual description and identifiable and retrievable via URIs.
- Wikipedia is popular as a reference, and we can thus expect its concept definitions to be agreed on by a broad readership. For instance, we can show that Wikipedia users have actively maintained at least 50 percent of all Wikipedia entries over at least 415 days, and can thus assume that substantial discourse has challenged and ultimately supported them (in the sense of Karl Popper, who has told us that trust in the truth of a theory can only be gained by continuously exposing it to falsification attempts).
- Wiki technology imposes only minimal requirements on users and is likely the simplest way to create a persistent URI with an informal description — anybody can add a URI for a needed concept at anytime.
- Researchers have made major efforts to mine Wikipedia content for formal semantics.[6–9] These approaches' impact will depend on Wikipedia URIs' ontological nature and conceptual stability, which no one has yet analyzed.

In general, multimedia components help clarify informal concept definitions in any vocabulary or lightweight ontology. Also, it's beneficial if a concept's definition isn't separated from the discussion that led to shaping its meaning because the discourse is important to that definition. In many disciplines, especially philosophy and the arts, you can understand a term's meaning only by knowing the historical debates that shaped it.

Challenges

When using wiki entries as conceptual elements, we face several research challenges.

Conceptual consistency over time. Anyone can modify wiki entries, and no explicit agreement exists between the user who created a new entry and those who modify it later. The concept a URI represents could change substantially over time, which would invalidate existing annotations — Wikipedia doesn't maintain a formal representation of the semantic differences between two versions (although ontology evolution research has suggested this for ontology engineering[10]), so wiki URIs could be unsuitable as authoritative identifiers.

A typical change is the introduction of *disambiguation pages*, which happens when the community realizes that the same label is a homonym and can be used with a different meaning in another context. In such cases, wiki users can turn the original page into a disambiguation page, which summarizes individual links for the context-specific entries. Thus, we must analyze whether Wikipedia's unsupervised, collaborative editing process can produce reliable identifiers for conceptual entities.

Multiple URI uses. One URI can denote multiple things; in particular, the URI for a retrievable document on the Web can identify either the entity the document defines or the document itself.[11] You could argue, for example, about whether http://en.wikipedia.org/wiki/Let_it_be denotes either this specific Wikipedia entry as a resource or the respective Beatles album. We must be able to distinguish both things because we might want to represent statements about each (that is, someone commenting on the Wikipedia entry versus someone referring to the album).

David Booth has already discussed two approaches to handling this problem[11] — either introduce different names for the two types of entities or make the context of the URI's usage explicit. Although the issue is nontrivial on a general Semantic Web level, we can agree on conventions for handling this issue in knowledge-management applications. If Wikipedia resources aren't needed as subjects of statements in an application, a given community can simply agree to view each Wikipedia entry as the entity that an average layman associates with this description. In our example, the URI would reflect the Beatles album, not the Wikipedia description of that album. This is a proposed social convention and is, of course, debatable, but it's reasonable in the context of our approach.

If such agreement isn't possible, we can solve the problem by using a different base URI for all Wikipedia entries when they're intended as identifiers for conceptual entities and not the documents themselves. However, the choice between those two approaches doesn't affect the findings we present later.

Lack of URIs for relationships and attributes. In general, a relevant community can deliberately agree on a wiki entry's content type and covered domain — for example, a wiki page could denote concepts, relations, or instances alike. Although wiki packages alone can also be used to define URIs for properties, by social convention manifested in Wikipedia guidelines, Wikipedia doesn't contain such entries. So although it would be technically possible to define a Wikipedia entry http://en.wikipedia.org/wiki/isAFriendOf, such entries aren't desirable. Also, users can't formally attach domains and ranges to such entries. Thus, quite naturally, what we mostly find in Wikipedia are abstract concepts and ontologically significant instances.

In knowledge-management applications, we also need to define properties and relationships as conceptual elements. We can achieve this in at least three ways:

- We can use properties and relationships from existing ontologies and Web vocabularies — namely, Dublin Core elements[12,13] — in combination with Wikipedia entries.
- We can create complementing property and relationship ontologies in an engineering fashion.
- We can set up a complementing "properties and relations" wiki, in which we can define URIs for properties informally using text descriptions. (Note that properties in several popular Web vocabularies, such as FOAF and vCard, are also defined in natural language only).

We can use all three approaches in combination, and as of this writing, we've already implemented the second and third approach. To simplify this article, we restrict our example to using attributes from the Dublin Core vocabulary.

Biased scope and imbalanced content. Wikipedia contains entries with varying degrees of abstraction and intertemporal relevance. Thus, content distribution is less balanced than in well-crafted vocabularies. Apart from lacking entries for relationships and properties, numerous entries reflect living or historical people. However, because Wikipedia's size has no upper limit, this problem isn't very relevant given that entries irrelevant for a particular purpose do no harm. Later, we estimate the proportions of entries in terms of Proton top-level categories.

Redundancy and dispersion. In Wikipedia, users can easily create multiple entries for the same concept. This has no negative impact on precision, but could lower the information retrieval recall when we use Wikipedia URIs as a reference vocabulary for annotations. To consolidate wiki content, wikis contain mechanisms for merging entries via redirects. In the popular MediaWiki software, for example, a user can insert the string `#redirect [[PAGENAME]]` into a discontinued page's body. As we show later, most redirects relate synonyms to each other. We could translate such links into statements of equivalence or semantic proximity (such as `rdf:seeAlso` or `skos:related`) in a knowledge-management application.

Example

Let's see how we can use Wikipedia entries (outside of the original wiki) in combination with Dublin Core attributes as a vocabulary for knowledge management. We based our example on the social convention that the reused Wikipedia entries identify the entity or concept that an average layman associates with this description, not the Web resource itself. In this sense, http://en.wikipedia.org/wiki/John_Lennon refers to John Lennon himself, not to the Wikipedia entry about John Lennon.

The example in Figure 1a represents that

- John Lennon was a contributor to the Beatles album "Let It Be,"
- the title of this Beatles album is "Let It Be (Beatles Album),"
- John Lennon is related to John Lennon's discography, and
- we can describe John Lennon with "John Winston Ono Lennon was a singer, songwriter, poet and guitarist for the British rock band The Beatles."

Figure 1b shows the resulting RDF graph.

Evaluation

To assess our proposal's feasibility, we evaluated current Wikipedia content with regard to its quality as a vocabulary for annotating Web resources. First, we looked at whether the Wikipedia URIs changed in meaning over their lifespan. Then, we estimated the proportion and total amount of entries according to the Proton ontology's top-level module categories. This let us assess the current Wikipedia content's ontological nature.

We took a representative, random sample of n = 150 pages from the English version of Wikipedia on 11 January and retrieved all respective content

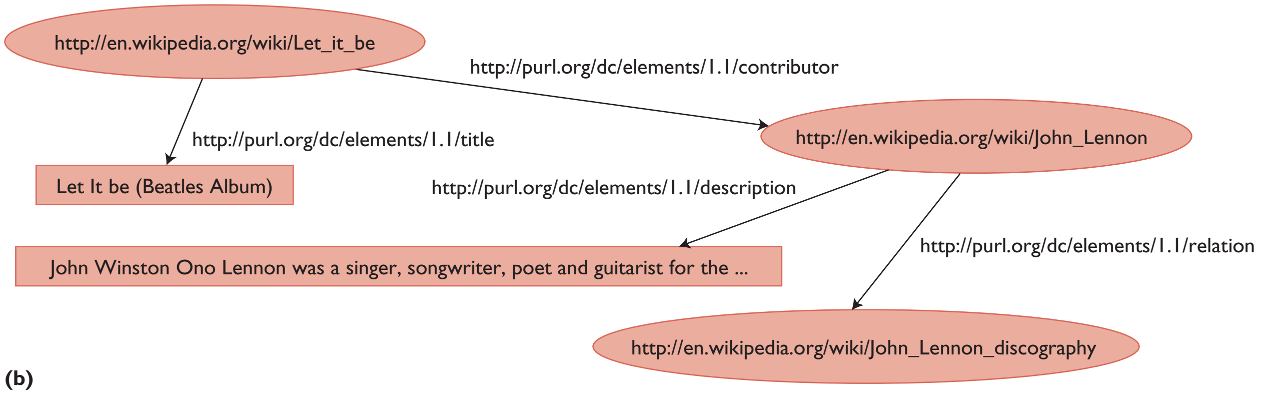
```
<?xml version=“1.0”?>
<!DOCTYPE rdf:RDF [<!ENTITY wiki “http://en.wikipedia.org/wiki/”>]>
<rdf:RDF xmlns:rdf=“http://www.w3.org/1999/02/22-rdf-syntax-ns#”
 xmlns:dc=“http://purl.org/dc/elements/1.1/”>

<rdf:Description rdf:about=“&wiki;Let_it_be”>
 <dc:title>Let It Be (Beatles Album)</dc:title>
 <dc:contributor rdf:resource=“&wiki;John_Lennon”/>
</rdf:Description>

<rdf:Description rdf:about=“&wiki;John_Lennon”>
 <dc:description> John Winston Ono Lennon was a singer, songwriter, poet and guitarist for the British rock band
      The Beatles.</dc:description>
 <dc:relation rdf:resource=“&wiki;John_Lennon_discography”/>
</rdf:Description>

</rdf:RDF>
```

(a)


http://en.wikipedia.org/wiki/Let_it_be
http://purl.org/dc/elements/1.1/contributor
http://purl.org/dc/elements/1.1/title
http://en.wikipedia.org/wiki/John_Lennon
Let It be (Beatles Album)
http://purl.org/dc/elements/1.1/description
http://purl.org/dc/elements/1.1/relation
John Winston Ono Lennon was a singer, songwriter, poet and guitarist for the ...
http://en.wikipedia.org/wiki/John_Lennon_discography

(b)

Figure 1. Describing John Lennon and his work using Wikipedia URIs. The example represents (a) several facts about John Lennon in RDF/XML and (b) the resulting RDF graph.

before 13 January 2007. This sample reflects approximately 0.01 percent of the population and returns sufficiently narrow 95 percent confidence intervals, as we show later. For gathering the sample, we used the MediaWiki software's “random page” functionality. Our sample doesn't include redirects, such as Wikipedia URIs that point to other pages or page fragments but contain no current description of their own. We can assume that the random number generator the MediaWiki software employs is sufficient for our analysis's purpose (for more information, see www.heppnetz.de/harvesting-wikipedia/).

Conceptual Reliability

To test whether the meanings the Wikipedia URIs denoted were stable despite continuous page edits, we used the following process. First, we evaluated whether the entity that the URI identified changed significantly between the first version and the current one — that is, we tested whether a layman would subsume the same objects under the initial version and the current version. We distinguished four cases that could occur:

- *Case 1 — no significant change in meaning.* The entry remains a stable concept from its very first version to the current one. Using the most recent version, we can correctly interpret all data annotated using the initial version. Also, all data annotated using the current version is valid with regard to the definition in the initial entry (although this is less relevant in practice because the retrievable resource on the Web consulted as a reference when expressing a query would be the current one).
- *Case 2 — minor change in meaning.* A minor change in meaning occurs, but we can still consistently interpret data annotated using the initial version. A typical case is an entry whose definition gets broader over time. Examples include entries that become disambiguation pages or new entries that get added to a disambiguation page. Because disambiguation pages typically represent “the set of entities often referred to as xyz,” a regular page changing into a disambiguation page falls into this category, as long as the initial entry is now among the more specialized, disambiguated ones listed.

- *Case 3 — major change in meaning.* A major change occurs in the entry's meaning; that is, we might interpret data annotated using the initial version incorrectly if we use the page's current definition.
- *Case 4 — deletion.* The initial entry is in our random sample but is deleted before we can evaluate its reliability and ontological nature. Although such deletions are disadvantageous for the user, given that he or she can no longer retrieve the respective resource, this doesn't automatically imply that the initial URI's meaning has changed. As with everything on the Web, no longer being a retrievable resource doesn't invalidate the meaning associated with a URI (just as the URI of a corporation that went out of business continues to denote that corporation). In general, quite a substantial number of delete operations occur in Wikipedia, but most happen very soon after an entry's creation. Such deletions follow clear rules and guidelines of relevance and appropriateness; we can see from the delete logs that most deletions occur to remove spam, advertisements or product placement, abuses of Wikipedia as private blogs, and so on.

Redirects — that is, wiki entries that turn into a redirect page — are a special type of change. Note that a redirection page is always recognizable as such (the MediaWiki software shows "Redirected from xyz" on the top of the page). Users are thus able to spot that the the page retrieved isn't the original content for that URI, but a related substitute. Most redirects (about 80 percent) just consolidate spelling variants or synonyms for the same page title. Redirects aren't included in our sample, but we provide additional results on the Web at www.heppnetz.de/harvesting-wikipedia/.

From the sample data of Wikipedia entries classified according to the four cases just specified, we next computed the proportion in the sample and both a Laplace and a Wilson point estimate[14] for the proportion in the population. In general, the Laplace approach returns a better estimate when the proportion in the sample is close to 0 percent or 100 percent. By multiplying the point estimate for the ratio with the population size (1,579,456), we can compute two alternative estimates for the total number of Wikipedia entries in each of the four cases.

We then computed the confidence intervals for these cases' proportions using the *adjusted Wald method* for a 95 percent confidence coefficient. Note that the popular textbook method for computing a confidence interval (the Wald method) is often unreliable, particularly when the proportions in the sample are close to the borders; an in-depth discussion of this problem is available elsewhere.[15,16]

Finally, for each entry, we also determined the following variables of the editing process and computed the mean, standard deviation, and quartiles Q1, Q2/median, Q3, and Q4/max:

- the entry's age as the difference between 13 January 2007 and the creation date;
- the discourse process's duration as the difference between the last edit and the creation date;
- the duration of unchanged existence as the difference between 13 January 2007 and the date of the last edit;
- the total number of versions; and
- the discourse process's average number of versions per day, which indicates the debate's intensity.

Our hypothesis is that despite Wikipedia entries' ongoing changes and uncontrolled editing, a stable community consensus exists about the meanings of most URIs.

Domain Focus

In addition to their conceptual reliability, we want to determine Wikipedia entries' ontological nature. For this, we manually classified all elements from the sample by Proton top-level categories. Then, we used similar techniques as in the previous section to estimate the total number of respective elements in the full Wikipedia population and estimated the number of conceptual entities in terms of the Proton categories.

For each entry in the sample, we determined the proper category in the Proton ontology's top-level module (http://proton.semanticweb.org/2005/04/protont), which has three main branches — *object*, *happening*, and *abstract* — and several subconcepts (such as person, group, and event), which we introduce later in the results section.

We then computed the proportion of each category in the sample, a Wilson point estimate[14] for the share and total number of respective entities in the population, and the confidence interval for the proportion based on a 95 percent confidence coefficient using the adjusted Wald method.[15]

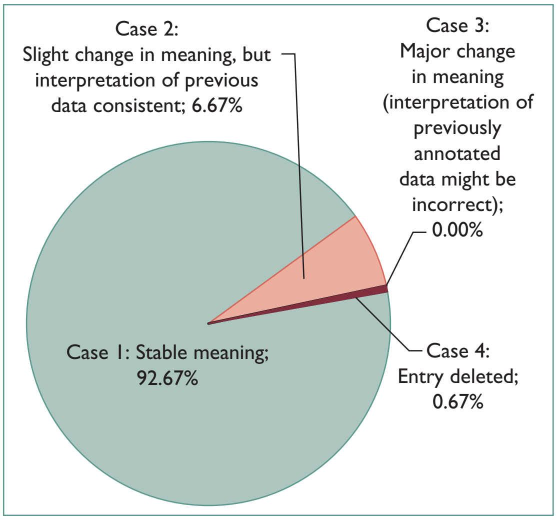


Figure 2. Reliability of Wikipedia URIs as identifiers. We can see that the majority of Wikipedia URIs refer to a stable conceptual element, and this despite the open editing process.

Results

The analysis of Wikipedia entries reveals two interesting results. First, despite the unsupervised, community-driven editing process, the conceptual entity associated with Wikipedia URIs rarely changes. Second, among the 1.5 million entries are very substantial amounts of concepts that are relevant for annotating Web resources, such as popular actors, research fields, cities, or universities.

Reliability of URIs as Identifiers

In the sample, 92.67 percent of the entries (n = 139) showed a completely stable meaning — that is, we can use them as identifiers for conceptual entities without any problems (Case 1). We found changes in the meaning in 10 elements (Case 2). These 10 entries (6.67 percent) have either always been disambiguation pages or became disambiguation pages. Those that had always been disambiguation pages had added new homonyms to the list, which accounts for the observed change in meaning, but doesn't really change the meaning as we've defined it. Most importantly, all the changes in meaning that we found broadened the concept's definition. Thus, when old data was annotated with the URI's original meaning in mind, the annotation remains valid with the current description for that URI. For example, a query expressed using the newly introduced or extended disambiguation page's current definition would still return only resources that belonged to the proper category (as defined by the current description for the URI). The only problem arising from Case 2 is that a search using the new disambiguated identifier wouldn't return elements referring to the old URI now being a disambiguation page. In other words, recall might be reduced, but precision isn't affected. (As an example, assume that there was an entry "xyz," initially referring to a city. Now, someone spots that there is also a play called "xyz" and turns the main entry into a disambiguation page that branches into "xyz (city)" and "xyz (play)". Thus, someone searching for resources annotated with "xyz (city)" won't find resources annotated in the past using the old URI of "xyz.")

We found no major changes in meaning (Case 3) in any of the 150 entries. One entry (0.67 percent) was deleted between determining the sample and capturing our sample data.

Note that although the number of actual deletions in Wikipedia is much higher, most of them are just the quick removal of inappropriate content. Also, a page's removal doesn't necessarily change the URI's meaning. In fact, the one deleted page in our sample refers to the band New London Fire. Wikipedia's deletion log shows that someone has attempted to create this page three times already (deleted on 9 June 2006; 6 October 2006; and 12 January 2007). We assume that the three attempts referred to the same band, in which case the deleted URI continues to refer to the same conceptual entity even if no page exists. Figure 2 shows the sample entries' conceptual reliability.

The findings are even more striking when we compute an estimate for the full population (see Table 1).

Both the Laplace point estimate (92.11 percent) and the Wilson point estimate (91.60 percent) indicate that more than 1.4 million URIs exist in Wikipedia that denote reliable conceptual entities. Table 1 gives the exact data. The 95 percent confidence interval for the stable proportion ranges from 87.22 percent to 95.98 percent. (Note that the computed lower limit of the confidence interval using the adjusted Wald method can be slightly below zero, even though common sense says it's zero. Moreover, computing the confidence interval for a case not found in the sample [zero occurrences] is a one-sided test.) We can estimate the number of URIs with a slight change in meaning as 114,353 (Laplace) or 122,387 (Wilson) Wikipedia entries, with a 95 percent confidence interval from 3.25 percent to 11.97 percent.

Because our sample has zero URIs with a major

Table 1. Reliability of Wikipedia URIs (sample and population estimates).

	Count	%	Laplace point estimate (%)	Population estimate	Wilson point estimate (%)	Population estimate	Lower limit (95%)	Upper limit (95%)
			Laplace method		Wilson method		Adjusted Wald method	
Case 1: Stable meaning	139	92.67	92.11	1,454,837	91.60	1,446,802	87.22	95.98
Case 2: Slight change in meaning, but interpretation of previous data consistent	10	6.67	7.24	114,353	7.75	122,387	3.52	11.97
Case 3: Major change in meaning (interpretation of previously annotated data might be incorrect)	0	0.00	0.66	10,424	0.89	13,992	–0.36†	2.13
Case 4: Entry deleted during our analysis	1	0.67	1.32	20,849	1.90	29,986	–0.26†	4.06

†These nonintuitive negative values are caused by the weaknesses of the adjusted Wald method when the proportion in the sample is close to zero.

Table 2. Amount of change and discourse in the Wikipedia sample.

	Age*	Duration of the discourse process**	Duration of unchanged existence†	Versions‡	Versions per day of the discourse process
Mean	574	508	66	37	0.147
Median	468	415	31	14	0.051
Standard deviation	460	460	95	79	0.422
Q1	199	135	8	7	0.026
Q2/median	468	415	31	14	0.051
Q3	850	758	88	32	0.112
Q4/max	1,992	1,957	541	720	3.667

* (13 Jan. 2007 — date of creation)

** (date of creation — date of last edit)

† (13 Jan. 2007 — date of last edit)

‡(as of 13 Jan. 2007)

change in meaning, we can assume that the total number of such problematic elements in the population is small. The Laplace point estimate (which is the most reliable point estimate for this case) is 0.66 percent of the population, meaning an estimated number of 10,424 such URIs. The confidence interval's upper limit is 2.13 percent.

Table 2 summarizes the data reflecting Wikipedia's age and the amount of modifications.

We can see from the median of 468 days for the age of all entries that almost half of them were added during the past 16 months (468/30 days), which is a strong indicator for continuous growth. As for the duration of an entry's unchanged existence, 25 percent of the entries have remained unchanged for no more than eight days; another 25 percent changed between 8 and 31 days before we took our sample. In the upper half of the sample (that is, values above the median), 25 percent of the entries changed between 31 and 88 days ago, and 25 percent remained unchanged for the past 88 to 541 days. Thus, entries' conceptual stability isn't due to the absence of change operations but occurs despite continuous editing from the community. We can conclude from this analysis that Wikipedia URIs are very reliable, authoritative identifiers for conceptual entities — likely, the biggest collection of this kind that's under the general public's full control.

Domain Focus and Wikipedia Entries' Ontological Nature

The analysis of Wikipedia entries' ontological nature shows that the majority of URIs in our sam-

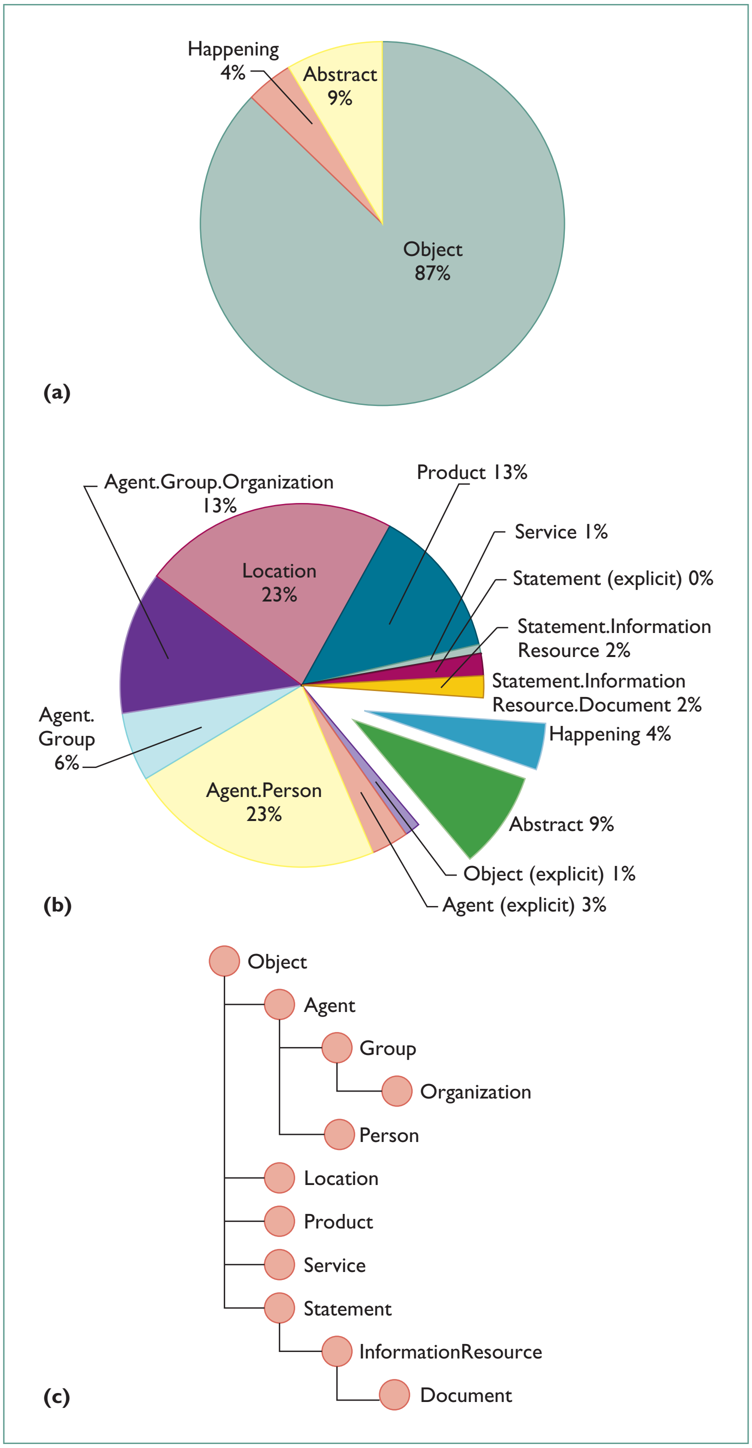


Figure 3. Wikipedia content by Proton top-module classes. (a) The general distribution of Wikipedia entries in terms of the three Proton ontology top-level classes abstract, happening, *and* object. *(b) A detailed breakdown of the* `protont.Object` *branch. We assigned each Wikipedia entry to the most specific subclass of* `protont.Object`. *For example, "Agent (explicit)" counts only those conceptual entities for which no more specific subclass of* `protont.Object.Agent` *exists. (c) The class hierarchy of the* `protont.Object` *branch.*

ple (87 percent) denote instances or subconcepts to the Proton top-level category *object*. This is defined as "entities that could be claimed to exist" (see http://proton.semanticweb.org). Nine percent are some sort of *abstract*, and 4 percent are classified as a *happening*. Figure 3a illustrates the proportion of entries in each main Proton category.

The breakdown of Wikipedia entries that fall into the `protont.Object` branch is very interesting. Figure 3b shows the proportions in the sample. (For statistical reasons, the point estimates for the population don't necessarily add up to 100 percent, which is why we based Figure 3 on the *sample* proportions and not on the population estimates from Table 3.) We always assigned each Wikipedia entry to the most specific subclass of `protont.Object`. This means that *agent* here counts only those conceptual entities for which no more specific subclass of `protont.Object.Agent` exists. We can see that the majority of the URIs denote people (23 percent), locations (23 percent), organizations (13 percent), product types (13 percent), and groups (6 percent).

If we look at Table 3, we can see from the Wilson point estimate that Wikipedia contains URIs for a substantial amount of significant conceptual entities:

- 368,790 living or dead people;
- 368,790 locations (namely buildings, cities, districts, and regions);
- 225,055 product types and models (roughly eight times the size of eCl@ss [www.eclass-online.com] or UNSPSC [www.unspsc.org]);
- 214,788 organizations; and
- 112,121 groups.

We don't know of any other reliable vocabulary that can be used for annotations that provides such a large and broad set of URIs that identify anything from regional high schools to living or dead people of all professions.

Our analysis shows that for the vast majority of Wikipedia entries, a community consensus exists about the URIs' meaning from the very first to the most recent version. In other words, open communities seem able to achieve consensus about named conceptual entities as very lightweight ontological agreements in an unsupervised fash-

Table 3. Wikipedia entries in terms of `protont.Object` ***(sample and population estimates).***

Proton	Count	Proportion in the sample (%)	Wilson point estimate (%)	Population estimate	Lower limit (95%)	Upper limit (95%)
			Wilson method		Adjusted Wald method	
Object	**130**	**86.7**	**85.75**	**1,354,401**	**80.23**	**91.27**
Object (explicit)	2	1.3	2.55	40,253	0.06	5.04
Agent (explicit)	5	3.3	4.50	71,054	1.22	7.77
Agent.Person	34	22.7	23.35	368,790	16.66	30.03
Agent.Group	9	6.0	7.10	112,121	3.04	11.16
Agent.Group.Organization	19	12.7	13.60	214,788	8.18	19.02
Location	34	22.7	23.35	368,790	16.66	30.03
Product	20	13.3	14.25	225,055	8.73	19.77
Service*	1	0.7	1.90	29,986	–0.26	4.06
Statement (explicit)*†	0	0.0	0.89	13,992	–0.36	2.13
Statement.InformationResource	3	2.0	3.20	50,520	0.42	5.98
Statement.InformationResource.Document	3	2.0	3.20	50,520	0.42	5.98
Happening	**6**	**4.0**	**5.15**	**81,320**	**1.66**	**8.64**
Happening (explicit)*	1	0.7	1.90	29,986	–0.26	4.06
Event	5	3.3	4.50	71,054	1.22	7.77
Situation (explicit)*†	0	0.0	0.89	13,992	–0.36	2.13
Situation.JobPosition*†	0	0.0	0.89	13,992	–0.36	2.13
Situation.Role*†	0	0.0	0.89	13,992	–0.36	2.13
TimeInterval*†	0	0.0	0.89	13,992	–0.36	2.13
Abstract	**13**	**8.7**	**9.70**	**153,188**	**5.02**	**14.38**
Abstract (explicit)*	1	0.7	1.90	29,986	–0.26	4.06
Number*†	0	0.0	0.89	13,992	–0.36	2.13
ContactInformation*†	0	0.0	0.89	13,992	–0.36	2.13
Language*	1	0.7	1.90	29,986	–0.26	4.06
Topic	2	1.3	2.55	40,253	0.06	5.04
GeneralTerm	9	6.0	7.10	112,121	3.04	11.16

* If *p* is small, the adjusted Wald method might return proportions below 0%, despite that we know that the actual lower limit is 0%.

† Because the number of this type of element in the sample is zero, the confidence interval is one-sided. Thus, we must use the z-value for a one-sided case (ca. 1.64) instead of the one for two-sided cases (ca. 1.96).

ion and relying only on the known mechanisms of standard wiki software to prevent destructive changes. We assume that the ease of access and using complementing multimedia elements for conceptualizing an entry are important factors in this process.

The mean of change operations per day per entry in the sample is 0.147. Multiplied by the number of entries, we can approximate the total number of change operations as 232,180 per single day, or roughly 7 million per month. This is about three times as much user involvement as 14 months ago, when we estimated a total of 2.5 million change operations per month. Despite this vast unsupervised user interaction, only a few URIs suffer from a change in meaning, and only a negligible amount might suffer changes so serious that interpretations of old data would be invalid. To us, this strongly indicates that heavy user involvement, debate, and broad usage are important contributors to truly shared domain conceptualizations. In fact, it might be that the sheer mass of user involvement produces more commonly agreed-on concepts than the more careful, more elaborate domain conceptualization that a small elite creates. (In a sense, Wikipedia's continuous use is a form of "pragmatic semantic unification" or successful validation of compatible world views by joint action.[17])

Although the result is only a flat vocabulary, we can try to augment our approach with additional semantics extracted from the actual Wikipedia data – for instance, researchers have had successful results automatically matching Wikipedia entries to WordNet synsets.[18] Recently, several others have reported successful approaches to mining Wikipedia for semantic structures.[6,7] For such approaches, our work provides empirical support of the content's stability and ontological nature. Finally, some related experiments successfully extracted semantic relationships between Wikipedia categories, which we could also use to augment a vocabulary based on Wikipedia URIs with additional semantics.[9] (The "Related Work in Ontology Engineering and Wiki Research" sidebar discusses more work in this area.) These recent

Related Work in Ontology Engineering and Wiki Research

Work related to ours mainly falls into five categories.

Community-Driven Ontology Building

Significant literature exists about collaborative ontology engineering in general, such as Tadzebao and WebOnto.[1] Jie Bao and Vasant Honavar describe collaborative ontology building in analogy to wikis,[2] but don't borrow more from the wiki community than the name, and they use a rich ontology metamodel as their starting point. They don't elaborate on ontology building's community focus or address the advantage of adding multimedia elements' informal concept descriptions.

Augmenting Wikis with Semantic Web Technology

Platypus Wiki[3] was an early wiki augmented by Semantic Web approaches — namely, RDF; our work uses wikis to create vocabularies usable anywhere in the Semantic Web. We present an early version of our work elsewhere,[4] but this prototype aimed to deploy a modified wiki installation as an ontology engineering platform. We now believe that Wikipedia must be the starting point due to its numerous existing entries and community pickup. Some researchers have described wiki extensions[5] so that users can explicitly augment the informal content with typed links and other formalized elements.

Mining Wikipedia Semantics

In contrast to providing technical support for adding semantics to wiki content, interest is growing in mining the semantics of Wikipedia content to provide access at a semantic level to the knowledge embedded in its entries.[6,7] The work presented in the main text complements such work by providing evidence on Wikipedia content's ontological nature and conceptual reliability. Also, we show that Wikipedia URIs alone constitute a valuable vocabulary for annotating Web resources.

Wikipedia Content

Some authors have recently analyzed the semantic coverage of Wikipedia and, in particular, the collaborative process of its production.[8] At a content level, others discuss Wikipedia's reliability as an encyclopedia (see http://en.wikipedia.org/wiki/Criticism_of_Wikipedia) but address only whether all facts said about a topic are authoritative in detail, not whether the URIs represent consensual concepts.

Stability of Web URIs

The stability of URIs on the Web has been analyzed elsewhere[9]; however, a major difference with our work is that the author evaluates whether URIs remain retrievable over time, whereas we analyze whether they keep reflecting the same concept over time and between many human contributors. A URI might well maintain its meaning, even if it's temporarily or permanently unavailable. A URI reflecting a bankrupt enterprise, for example, could well continue to serve as a unique identifier for that company.

References

1. J. Domingue, "Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web," *Proc. 11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1998; http://kmi.open.ac.uk/~john/banff98-paper/domingue.pdf.
2. J. Bao and V. Honavar, "Collaborative Ontology Building with Wiki@nt — A Multi-Agent Based Ontology Building Environment," *Proc. 3rd Int'l Workshop on Evaluation of Ontology-based Tools* (EON 04), 2004; http://citeseer.ist.psu.edu/722905.html.
3. S.E. Campanini, P. Castagna, and R. Tazzoli, "Platypus Wiki: A Semantic Wiki Web," *Proc. 1st Italian Semantic Web Workshop Semantic Web Applications and Perspectives* (SWAP), 2004; http://semanticweb.deit.univpm.it/swap2004/cameraready/castagna.pdf.
4. M. Hepp, D. Bachlechner, and K. Siorpaes, "OntoWiki: Community-Driven Ontology Engineering and Ontology Usage Based on Wikis," *Proc. Int'l Symp. Wikis* (WikiSym 06), ACM Press, 2006, pp. 143–144.
5. M. Völkel et al., "Semantic Wikipedia," *Proc. 15th Int'l Conf. World Wide Web* (WWW 06), ACM Press, 2006, pp. 585–594.
6. F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," *Proc. World Wide Web Conf.* (WWW 07), ACM Press, 2007, pp. 697–706.
7. S. Auer and J. Lehmann, "What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content," *Proc. 4th European Semantic Web Conf.* (ESWC 07), LNCS 4519, Springer, 2007, pp. 503–517.
8. T. Holloway, M. Bozicevic, and K. Börner, "Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors," *Complexity*, vol. 12, no. 3, 2007, pp. 30–40.
9. D. Spinellis, "The Decay and Failures of Web References," *Comm. ACM*, vol. 46, no. 1, 2003, pp. 71–77.

efforts could be an important starting point for deriving taxonomic relations and even a grounding in Proton, thus extracting a true formal ontology from Wikipedia.

Acknowledgments

We thank Elias Kirche from Florida Gulf Coast University for many discussions on point estimates (and for numerous cups of coffee), and the anonymous reviewers for their detailed and helpful comments. The work presented in this article has been supported by the European Commission under the projects SUPER (FP6-026850) and MUSING (FP6-027097), and by the Austrian BMVIT/FFG under the FIT-IT project myOntology (grant no. 812515/9284). Martin Hepp also has support from a Young Researcher's Grant (Nachwuchsförderung 2005-2006) from the Leopold-Franzens-Universität Innsbruck.

References

1. N. Guarino, "Formal Ontology and Information Systems," *Proc. 1st Int'l Conf. Formal Ontology in Information Systems* (FOIS 98), IOS Press, 1998, pp. 3–15.
2. T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *Int'l J. Human-Computer Studies*, vol. 43, nos. 5/6, 1995, pp. 907–928.
3. D. Fensel, *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*, 2nd ed., Springer, 2004.
4. M. Hepp, "Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies," *IEEE Internet Computing*, vol. 11, no. 1, 2007, pp. 90–96.
5. T.R. Gruber, "Ontology of Folksonomy: A Mash-Up of Apples and Oranges," *Int'l J. Semantic Web and Information Systems*, vol. 3, no. 1, 2007, pp. 1–11.
6. F. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," *Proc. World Wide Web Conf.* (WWW 07), ACM Press, 2007, pp. 697–706.
7. S. Auer and J. Lehmann, "What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content," *Proc. 4th European Semantic Web Conf.* (ESWC 07), LNCS 4519, Springer, 2007, pp. 503–517.
8. M. Völkel et al., "Semantic Wikipedia," *Proc. 15th Int'l Conf. World Wide Web* (WWW 06), ACM Press, 2006, pp. 585–594.
9. S. Chernov et al., "Extracting Semantic Relationships between Wikipedia Categories," *Proc. 1st Int'l Workshop: SemWiki2006 – From Wiki to Semantics* (SemWiki 06), CEUR WS Proc., vol. 206, 2006; http://CEUR-WS.org/Vol-206/.
10. N.F. Noy and M. Klein, "Ontology Evolution: Not the Same as Schema Evolution," *Knowledge and Information Systems*, vol. 6, no. 4, 2004, pp. 428–440.
11. D. Booth, "Four Uses of a URL: Name, Concept, Web Location and Document Instance," 2007; www.w3.org/2002/11/dbooth-names/dbooth-names_clean.htm.
12. Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, version 1.1: Reference Description," 2005; http://dublincore.org/documents/dces/.
13. Dublin Core Metadata Initiative, "DCMI Metadata Terms," 2005; http://dublincore.org/documents/dcmi-terms/.
14. V. Chew, "Point Estimation of the Parameter of the Binomial Distribution," *The Am. Statistician*, vol. 25, no. 5, 1971, pp. 47–50.
15. L.D. Brown, T.T. Cai, and A. DasGupta, "Interval Estimation for A Binomial Proportion," *Statistical Science*, vol. 16, no. 2, 2001, pp. 101–131.
16. E.B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *J. Am. Statistical Assoc.*, vol. 22, no. 158, 1927, pp. 209–212.
17. C. Petrie, "Pragmatic Semantic Unification," *IEEE Internet Computing*, vol. 9, no. 5, 2005, pp. 96–C3.
18. M. Ruiz-Casado, E. Alfonseca, and P. Castells, "Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets," *Proc. 3rd Atlantic Web Intelligence Conf.* (AWIC 05), LNCS, Springer, 2005, pp. 380–386.

Martin Hepp is a professor of computer science at the University of Innsbruck, Austria, where he heads the Semantics in Business Information Systems research unit. He created eClassOWL, the first industry-strength ontology for products and services, and is currently working on using Semantic Web technology for business applications — namely, e-commerce and ERP systems management. Hepp has a master's degree in business management and business information systems and a PhD in business information systems from the University of Würzburg, Germany. He is a member of the IEEE Computer Society. Contact him at mhepp@computer.org; www.heppnetz.de.

Katharina Siorpaes is a PhD student in computer science at the University of Innsbruck, Austria. As part of her PhD research, she's analyzing the usage of Wiki technology for the community-driven design and maintenance of ontologies, in particular for e-commerce applications. Siorpaes has a BSc and an MSc in computer science from the University of Innsbruck. Contact her at katharina.siorpaes@deri.at.

Daniel Bachlechner is a master's student in computer science and international business administration at the University of Innsbruck, Austria. His main research interest is in analyzing the business value of semantic technology. Bachlechner has a BSc in computer science from the University of Innsbruck. Contact him at daniel.bachlechner@deri.at.