Introducing new features to Wikipedia

– Case studies for Web Science –

Mathias Schindler
Wikimedia Deutschland e.V.
Deutsche Nationalbibliothek
Germany
mathias.schindler@wikimedia.de

Denny Vrandečić
Institute AIFB
Universität Karlsruhe(TH)
Germany
denny@aifb.uni-karlsruhe.de

ABSTRACT

Wikipedia is a free web-based encyclopedia. It is written in collaboration by hundred of thousands of contributors [2]. It runs on the Open Source MediaWiki wiki engine. Introducing new features to Wikipedia is not just a technical question, but a complex socio-technical process. Previous introductions of new features (the category system in 2004, parser functions in 2006, and flagged revisions in 2008) are described and analyzed. Based on these experiences, the design of a new feature (creating semantic annotations) is given and discussed. The interaction between the technical features and the community is shown to be an instantiation of the Web Science research cycle, thus testing the cycle as a methodological tool for web science.

1. INTRODUCTION

Wikipedia is a free web-based encyclopedia. It is written in collaboration by hundred of thousands of contributors [2]. Today, Wikipedia is available in more than 250 languages offering more than 10 million articles. For some of these languages, Wikipedia is not only a free encyclopedia, but also the only encyclopedia available in this language. Currently it is among the ten most visited web sites in the world, sporting more than 50 thousand hits per second in the peak time.

The introduction of new technical features to Wikipedia has always turned out to be a complex socio-technical process. Since both the technical implementation and the content development of Wikipedia are done in free and open projects that both offer their complete change logs,[1] the co-development of the technical and social aspects can be traced and analyzed.

Section 2 offers a short overview of the web science process, before using it to describe the introduction of new features to Wikipedia in four case studies: the category system, introduced in 2004 (Section 3.1); parser functions, introduced in 2006 (Section 3.2); flagged revisions, introduced in 2008 (Section 3.3); and semantic annotation, currently being proposed (Section 4). We close the paper in Section 5 with conclusions drawn from the case studies.

2. WEB SCIENCE PROCESS

The web science process [4] is a model to describe the evolution of the web and of systems on the web. In order to resolve issues, engineers creatively come up with new ideas. These are then implemented. Since the web is an inherently social infrastructure, the technical implementation of an idea will necessarily involve social aspects. If done well, the socio-technical implementation of the idea may resolve the original issue on a micro-level. But due to the size and the complexity of the web, the solution will have novel and unexpected implications on the macro-level. These in turn have to be analyzed in order to identify new issues (i.e. situations that do not conform to certain values), thus starting the web science process anew. A detailed description of the steps of the process is given in [4]. Figure 1 gives an graphical overview of the model [3].

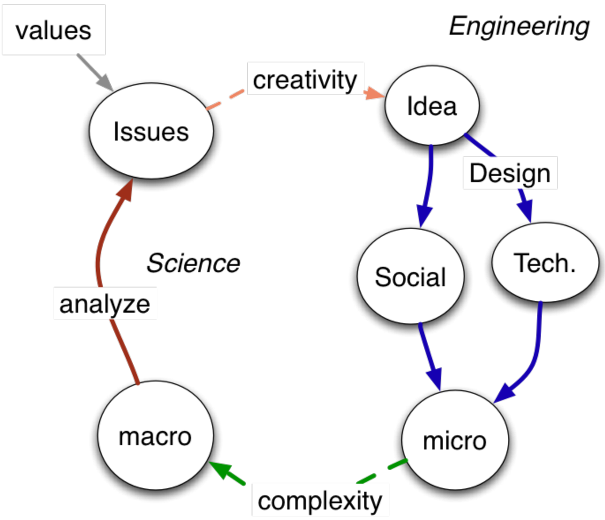

Figure 1: The web science process as presented by Tim Berners-Lee in his Keynote *The two Magics of Web Science* at the WWW2007, Banff, Canada. The two magics are creativity and complexity, since both are not well understood yet. Source: `http://www.w3.org/2007/Talks/0509-www-keynote-tbl/#(11)`

The authors already identify a number of examples for the web science process. In this paper we describe further examples from the history of Wikipedia, that can be fully traced and analyzed due to the availability of the data about the Wikipedia project and its history.

3. PREVIOUS EXPERIENCES

This section describes three case studies of previously introduced features to the Wikipedia system. Section 3.4 notes some other examples that may grant further studies.

[1] MediaWiki's SVN is accessible at `http://svn.wikmedia.org`, the database dumps of Wikipedia's complete log history can be found at `http://download.wikipedia.org`

3.1 Category system

Following the initial growth of Wikipedia, features to improve exploring and navigating the content became necessary. Early versions of the MediaWiki software offered basically only the manually created links, a backlinks feature, and a full text search in order to discover the content of the encyclopedia. In many other wikis the backlinks feature (that displays all pages that link to a specific page) was used to implement a rudimentary tagging system: on creating a link to a page like "Greece-related topic" on all pages that discuss topics related to Greece, the list of backlinks on the page "Greece-related topic" will be a list of all pages on that topic. This approach has a number of disadvantages: the links to the topic page will behave like any other link (i.e. display in the text), the linked page itself will behave like a normal article (i.e. it will be a normal page, and displaying the backlinks will require a second click on the appropriate link in the toolbar), and normal articles like "Greece" can not be used as category pages (since all normal links would also appear in the list, not just those that were introduced for categorizing: Albania would be in the list of pages linking to Greece, since it is would link to it, being a bordering country). Even though this is the usual procedure in many other wikis, Wikipedia often avoided such wiki-specific idiosyncrasies, e.g. the omission of camel-case syntax in favor of free links (leading to Wikipedia often being described as a rather atypical wiki).

The category system was introduced to MediaWiki in 2004 in order to solve that problem. Each article could be put into an arbitrary number of categories that are identified by freely chosen names. Adding a page to the category "Greece" is done by adding `[[Category:Greece]]` to the page. Category links themselves are not displayed in the article text (but rather in separate visual elements on the rendering of a page). The category page is a page of its own in the newly introduced category namespace, thus separating it cleanly from the article space. Category pages were programmed in such a way to display the list of all pages categorized with this category. The new category system resolved all the identified technical issues of using backlinks for categorizing articles.

The category system was activated in all language editions at the same time without consulting the Wikipedia communities of the different languages. On most languages it was quickly applied to categorize a majority of the existing pages. Soon new issues were discovered, namely how to organize categories themselves. This lead to further new features like the category tree extensions, which renders a tree of subcategories on the page of a category etc. The introduction and application of the categories were heatedly debated in some language communities. Restrictions on the usage of categories were introduced and enforced solely by the community and with manual effort. On the German Wikipedia, the community decided on a moratorium on category usage, in order to first define some guidelines to their application. The moratorium, and the enforcement of the guidelines, were done fully manually by the community members.

Back then the power to develop, enable and disable new features lay basically with the software developers. The communities around the different language Wikipedias had no option but to accept these changes and deal with the aftermath. Technical details – like the fact that redirects on categories did not mean that a contributor could use the redirecting category as a synonym for categorization – had enormous social impact, since they formed the patterns of interaction with the software. An analysis of the category system is given in [6]

(a)
```
Greece is {{Country | Continent=Europe
| Capital=Athens}}
```

(b)
```
a country in {{{Continent}}}.  It's
capital is {{{Capital}}}.

[[Category:Country]]
```

(c)
```
Greece is a country in Europe.  It's
capital is Athens.

[[Category:Country]]
```

Figure 2: Usage of templates. (a) Source of a page about Greece calling the country template; (b) Source of the country template; (c) Text of the page about Greece after template-expansion.

3.2 Parser functions

Templates are used to include the text of another page (usually in the separate Template namespace) at the place were the template is being called. This allows for a higher consistency within the wiki, since some elements will have to be written only once and reused in several articles. Templates may also feature parameters, i.e. parametrized template calls will insert the parameter's values in the replacing text. See Figure 2 for an example. Templates are widely used in Wikipedia, e.g. the English Wikipedia offers more than 200,000 templates.

In 2006 some Wikipedians discovered that through an intricate and complicated interplay of templating features and CSS they could create conditional wiki text, i.e. text that was displayed if a template parameter had a specific value. This included repeated calls of templates within templates, which bogged down the performance of the whole system. The developers faced the choice of either disallowing the spreading of an obviously desired feature by detecting such usage and explicitly disallowing it within the software, or offer an efficient alternative. The latter was done by Tim Starling, who in April 2006 announced the introduction of parser functions,[2] i.e. wiki text that calls functions implemented in the underlying software.

At first, only conditional text and the computation of simple mathematical expressions was implemented, but this already increased the possibilities for wiki editors enormously. With time further parser functions were introduced, finally leading to a framework that allowed the simple writing of extension function to add arbitrary functionalities, like e.g. geo-coding services or widgets. This time the developers were clearly reacting to the demand of the community, being forced either to fight the solution of the issue that the community had (i.e. conditional text), or offer an improved technical implementation to replace the previous practice and achieve an overall better performance.

3.3 Flagged revisions

The most obvious issue with a wiki-based encyclopedia is the fact that since anyone can edit all pages, defaced articles may be en-

[2] *"In response to a campaign by users of the English Wikipedia to harass developers by introducing increasingly ugly and inefficient meta-templates to popular pages, I've caved in and written a few reasonably efficient parser functions. There are two conditional functions and a mathematical expression function."* – Tim Starling, Wed, Apr 5, 2006

countered. The first approach towards dealing with this problem was to protect some articles that tended to be defaced often (e.g. George W. Bush, Abortion, or the Main Page). Protection means that the pages were only editable by users that have administrator rights. Naturally, this approach decreases "wikiness". As a result, "semi protection" was introduced, making articles editable by all registered users. Preliminary research indicates that semi protection greatly reduces risk of vandalism, especially of "drive-by" vandalism; it does not, however, severely affect the collaborative process.

Flagged revisions enable a second layer to store the ad-hoc validation of pages. After a trial phase, it is now a permanent feature in the German language Wikipedia. In February 2009, all articles in this language edition were assigned at least one flagged revision. The remaining work is to keep up with edits by unregistered or new users. Any Wikipedia author who is a member of the "editor" group can mark a revision of an article to be compliant with a set of conditions on content quality. The current condition for a revision to be flagged is the "lack of obvious forms of vandalism", a deliberately low threshold. Unregistered visitors to Wikipedia (the vast majority of the visits to Wikipedia) will be shown the latest flagged revision instead of the most recent article version.

Both prior to implementation and after the initial start of this feature, a lengthy debate was held within the author community, resulting in a straw poll. The preliminary result was a rather strong mandate to keep the feature enabled. Other language projects were given the choice to use this extension (the discussion on the English Wikipedia is being followed by newspapers). Developers and the community were tightly working together to realize this solution. This feature is expected to make a strong impression on the casual Wikipedia readers by ensuring that no spam or profanity is displayed to them. Based on the exhaustive debate it is expected that the feature will indeed achieve to effectively confine vandalism, and as of now the feature seems to function as intented.

3.4 Further examples

This section points out to some other features that we consider worthwhile for further investigation due to their specifics differentiating them from the previous examples.

Some early extensions may be surprising in their scope: the Hierowiki extension allowed to use Hieroglyphs within wiki text. This was created and advocated by a small user group, but due to the early state of the overall project it was swiftly implemented and integrated. Later the functionality needed to be refactored out of the core, without affecting the usage in the Wikipedia content.

The Cite extension created by core developer Magnus Manske allows to add citation information to an article and render them appropriately. The Cite extension introduces new, XML-inspired syntax to the wikitext, which is rather inconsistent with the rest of the wiki-syntax. The community seemed to be not too bothered by that inconsistent syntax, but rather quickly adopted (and extended) the possibilities of the new extension.

The MediaWiki software is further extending the possibilities to enable personalized for logged in users. Within the wiki, users may add new Javascript and CSS commands that will be included in the served HTML pages for the user. Some users offer Javascript snippets that can be integrated into the user's own Javascript extension page. This framework allows for almost arbitrary adaptations of the individual Wikipedia experience, e.g. by adding a more powerful editor, or by hiding parts of the page output. The Widget extension takes this to a new step, allowing less technically inclined users to access some of these personalization features as well. This decouples the user's individual experience from the community and allows them to individually enhance Wikipedia with new features.

A recurring topic among developers and active users has been the conversion of wiki text or rendered HTML output to PDF. Early attempts were mostly based on manual work (like the WikiReader project), involving OpenOffice templates and its PDF output. These projects triggered several pieces of software that implemented a Medawiki to PDF converter or a different way to create more visually appealing Wikipedia articles for print. In 2008, the Wikimedia Foundation announced a partnership with PediaPress GmbH, where the commercial partner would create and maintain the software to create the output. This shows that also modules developed outside of the MediaWiki Open Source developer community can become integrated into Wikipedia.

4. SEMANTIC ANNOTATIONS

The Semantic MediaWiki project [5] allows to add structured data to wiki pages and exploit these both within the wiki through querying and browsing, and outside of the wiki by using the RDF export of the annotation structure. In order to create the structured data, an extension of the MediaWiki syntax is employed to allow for typed links and further annotation. On the page for Greece the wiki text `[[Capital::Athens]]` would be interpreted as the triple "Greece - Capital - Athens".

It was soon noted that these annotations make the wiki text harder to read and may confuse and discourage novice contributors. To solve this, editors often moved the annotations from the text to templates (see Section 3.2). For example, on the page of Greece the Country template may be called, using the parameter "Capital":

```
{{Country|Capital=Athens}}
```

The template can now create a consistent layout for country pages, e.g. by including an infobox displaying the capital. In Semantic MediaWiki, the template can also take care of annotating the country article with the capital, thus moving the additional annotation syntax from the article to the template. This shifts the burden from the casual Wikipedia editor to people who are comfortable with editing the already rather complex template syntax. Note that this is not a technically enforced trade-off: the annotation may be either in the article or in the template. It is up to the community to choose and enforce how annotated links should be used.

Instead of annotating links, the community may also choose to put the annotations at the bottom of the page, similar to the way category annotations are used today. So, articles would still include the above annotation, but no link would be displayed at that place. This decouples the annotations from the text. An example is given in Figure 3. We see that in option (b) the actual links in the text are annotated, whereas option (c) has all annotations at the end. The decoupling has a different trade-off: having the links annotated in the text increases consistency (since changing the capital of Greece from Athens to Sparta would change both the text and the annotation); having the annotated links at the end increases readability. Again, this is a social choice – the technology allows for both options: simplifying editing or increasing consistency. But this may only be the micro-view on the issue: simplified editing could in

(a)

```
Greece is a [[Europe]]an country.  It's
capital is [[Athens]].

[[Category:Country]]
```

(b)

```
Greece is a [[Continent::Europe]]an
country.  It's capital is
[[Capital::Athens]].

[[Category:Country]]
```

(c)

```
Greece is a [[Europe]]an country.  It's
capital is [[Athens]].

[[Category:Country]]
[[Continent::Europe]]
[[Capital::Athens]]
```

Figure 3: Source of a page about Greece in (a) plain MediaWiki, (b) Semantic MediaWiki, (c) Semantic MediaWiki using category-style annotations.

turn lead to increased community participation which may eventually lead to more consistency due to an increased number of eyeballs. It is safe to say that predicting the socio-technical impact of such a choice is rather hard.

An alternative approach to extract the data from Wikipedia is offered by the DBpedia project [1]. DBpedia extracts the data from the existing templates, not changing the technical component of Wikipedia at all. Instead it creates a strong incentive for standardizing and controlling the usage of templates with much more scrutiny. With DBpedia becoming more successful this may lead to more pressure to change the way templates are used in order to achieve a better information extraction – thus pushing for a much stronger change in the social aspect than the bigger technical change of Semantic MediaWiki does. This again is a trade-off: bigger technical changes by implementing Semantic MediaWiki on the Wikipedia servers, or eventually leading to larger changes in the way the community applies templates.

5. OUTLOOK AND CONCLUSIONS

Due to its open nature and the abundant and freely available data about the project and its history, Wikipedia is frequently used as an object of study. However, both the conducted research and the examples of feature implementations suggest that the better the complex interaction of technical progress and social adaption is understood, the more likely a new technology can be successfully introduced to Wikipedia. This interaction may be regarded as the *third magic* of Web science, adding to the *two magics* already described by Tim Berners-Lee (see Fig. 1).

We are convinced that we can learn a number of lessons from analyzing the introduction of new features in Wikipedia. The preliminary survey in this paper indicates for example the following lessons:

- The lack of some technical features may lead to the misuse of other existing features in order to simulate the desired functionality (as exemplified by the use of backlinks for categories, Section 3.1)
- Common usage practices of the community may necessitate changes in the technical platform in order to meet technical constraints without changing the practice of the community too much (as seen for parser functions, see Section 3.2)
- The introduction of a feature may lead to a huge perceived change in the mission and scope of the whole system (a challenge posed by the flagged revisions, see Section 3.3)
- In a socio-technical system, imperfections in the underlying technical platform can be compensated by the community (rf. to the unusual syntax of the Cite extension, Section 3.4)
- A priori design of new features must not only regard and resolve the technical challenges, but also consider possible social implications. Our current plan for the eventual implementation of Semantic MediaWiki in Wikipedia is to openly discuss the different alternatives with the community, and let them decide on the actual implementation.

This paper has argued that Wikipedia can be used to evaluate and validate the proposed web science process model. We have used the descriptive power of the model for both the a posteriori analysis and the a priori design of new features for Wikipedia.

Acknowledgments

Work presented in this paper has been partially funded by the European Union under EU IST FP7 project ACTIVE.[3]

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. *6th Int. Semantic Web Conference, Busan, Korea, Nov*, 2007.

[2] P. Ayers, C. Matthews, and B. Yates. *How Wikipedia works*. No Starch Press, San Francisco, CA, Oct. 2008.

[3] T. Berners-Lee. The two magics of web science. Keynote at the WWW2007 conference in Banff, Canada.

[4] J. Hendler, N. Shadbolt, W. Hall, T. Berners-Lee, and D. Weitzner. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, 2008.

[5] M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller, and R. Studer. Semantic Wikipedia. *Journal of Web Semantics*, 5:251–261, Sept. 2007.

[6] J. Voss. Collaborative thesaurus tagging the Wikipedia way. *The Computing Research Repository*, abs/cs/0604036, 2006.

[3] http://www.active-project.eu)