elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Operation of a Large Scale, General Purpose Wiki Website

Experience from susning.nu's First Nine Months in Service

Lars Aronsson

Aronsson Datateknik,
Teknikringen 1 e, 58330 Linköping, Sweden
(phone) +46-70-7891609, lars@aronsson.se

Abstract. A Wiki website is a hypertext on steroids. Any user can create or edit any page on the site using a simple web browser, and all information processing is done on the server side. Wiki sites are powerful tools for collaboration in closed work groups, but can also be used for the general public on the open Internet. This paper summarizes the experience from the first nine months of operation of Sweden's biggest Wiki website susning.nu, including its usefulness in non-profit and commercial applications, in hobby and professional, projects, its social and legal aspects, its relation to geographic information systems, subject information gateways, the establishment of a controlled vocabulary, and its implications on learning, free speech, the price of information, licensing, and copyright. Relevant comparisons to similar projects in other countries are also presented.

1 Introduction

The first section gives an introduction to the people and groups that use wiki websites. The next two sections give an introduction to the technology itself and some legal aspects. The fourth section contains a description of my own Swedish wiki website, susing.nu.

2 The Wiki Movement

After they were slashdotted on March 22, 2001, members of the Seattle Wireless group reported that the front page of their website had been damaged. Apparently, this wireless community networking group in Seattle, Washington (www.seattlewireless.net), had a link on their website that allowed anybody to edit the page, and someone had done so. However, it seemed they had an earlier version saved and could easily restore it. When I first heard about this, I was surprised that they could be so naive to believe that their website would be left alone with that feature open to the public.

Later that spring, I learned about Wikipedia, C2.com, and several other websites that were using a similar technology. It is called WikiWikiWeb or WikiWiki or Wiki, and its core principle is that every page of the website can be edited by anybody at any time. If the World Wide Web was a radical change in the way we publish information, Wiki seems to mean total anarchy. So why would anybody ...?

The first wiki website is said to have been launched already in 1994. The term wiki was coined by Ward Cunningham, who is acknowledged as the inventor of this concept. He lives in Portland, Oregon, and is a software developer and a consulting project manager to large development projects. To manage the complexity of such projects, he has become a pioneer in new software methodologies such as Object-Oriented Programming, Design Patterns and Extreme Programming.

One critical aspect of software development is documentation, not only producing a user's manual when the system is ready, but maintaining technical specifications for use by the developers during the project. This technical documentation has to change over time, and needs to be updated by many people. Traditionally, this need is filled by storing word processing documents on a shared file server, but this method has several drawbacks: Word processors are made for producing printed documents. Hypertext links might not be supported. Revision control and the ability to trace a document's history might not be an integral part of the system. The process for updating and approving a new version of a document can be slow, resulting in documentation that is constantly out of date. These drawbacks can make developers resort to e-mail rather than written documentation as a source of knowledge about the system under development–a sure way to lose grasp of complexity. A working documentation system must be fast, powerful, easy to use, and highly automated, otherwise developers will avoid using it.

What Ward did, was to design a small server-side CGI script on the project group's intranet webserver, that allowed any member of the group to make instant updates to the online documentation. Each change was logged and could be reviewed and compared to previous versions by the group members, just by using their web browsers. This turned out to be a very fast way to create and maintain the technical documentation, and Ward chose to call it wiki-wiki, a Hawaiian word that means fast or quick. Later he opened a public wiki on his company's external website, c2.com, for discussing matters of software engineering and methodology, called the Portland Pattern Repository. Still in early 2001, this was the world's biggest wiki.

Since then, wiki has caught on as a community building tool for various technology-oriented interest groups. One of the biggest wiki websites is the FoxPro Wiki (`http://fox.wikis.com/`) for users of the FoxPro database software, having several thousand pages. Most wiki websites, however, have just a handful or a few hundred pages. Many of the wireless community networking groups that popped up in 2000 and early 2001 also run wiki websites, as do many of the open source projects at SourceForge.

A good starting point for discussions about wiki technology and community building is the MeatballWiki website, `http://www.usemod.com/cgi-bin/mb.pl`

The Wikipedia, an invention of Larry Sanger in January 2001, marks the beginning of a new era, using wiki as a tool for building a general-purpose, free, and open encyclopedia.

The dream of collecting all human knowledge within arm's reach can be traced to ancient libraries of Alexandria and Pergamom, Denis Diderot (Encyclopédie, 18th century), Vannevar Bush (Memex, 1945), and Ted Nelson (Project Xanadu). In the last decade, such dreams have focused around the Internet, with the Tim Berners-Lee's World Wide Web itself being the most successful and least constrained example. Many projects with an explicit goal of creating a web-based, open source encyclopedia have started, but all have failed to attract any substantial interest.

Among these less successful web-based encyclopedia projects is Nupedia, started by Jimmy Wales, owner of the Bomis web portal, living in San Diego, California. Jimmy hired Larry Sanger as editor in chief for Nupedia, with the role of coordinating voluntary contributions of article submission and review. With very high quality ambitions, it turned out that not many articles were produced, and it would take long before the Nupedia could be a useful source of knowledge. This is when Larry, in January 2001, took the initiative to start a wiki website, as a lightweight parallel project, where anybody could edit articles.

Larry Sanger named this wiki the Wikipedia. He used the UseModWiki software (`http://www.usemod.com/`) written by Clifford Adams. In its first year of existence, Wikipedia has become the world's biggest wiki website by far, having more than 30,000 pages or articles, and the first really successful attempt to create a free, web-based encyclopedia.

Wikipedia started out as a single wiki website in English, but the success of this initiative soon led to a demand of parallel versions in other languages, and Jimmy Wales has set up independent stub Wikipedias for a large number of languages. A handful of these have attracted substantial user groups, including Spanish, German, Polish, Esperanto, French, and Dutch.

In early 2002, Jimmy Wales had to cut Larry Sanger from his staff, and the project is now entirely based on voluntary contributions. New software, that has been written by Wikipedia volunteers, scales better to the size of the Wikipedia, and the project is thriving. Also in early 2002, a group of contributors to the Spanish Wikipedia decided to break out and set up a server of their own at the University of Sevilla. Their "Enciclopedia Libre" now has 9,000 pages, while the Spanish Wikipedia has stopped at the 2000 pages it had at the time of the fork. More about size comparisons later.

3 Wiki Technology

A wiki is implemented as a website component, a CGI script, or any similar server-side scripting technology. Several open source, free software implementations exist. This script manages a set of small documents, known as wiki pages. Each wiki page has a unique name, that describes the subject matter of that page. Each wiki page can be displayed as a web page, using its name as part of the URL. This URL will typically be in CGI syntax containing the name of the CGI script, e.g. `http://myhost.com/cgi-bin/wiki.php?view=Library`

The wiki page itself is written in plain text and stored either on file or in a database. When a browser requests a page, the wiki script translates this plain text into HTML that becomes part of the returned web page. In addition to this, the web page also contains a header showing the name of the page, a navigation menu, and some links that are specific to the displayed page. The most important of these links – and this is the essence of the wiki concept – says "edit this page", e.g. `http://myhost.com/cgi-bin/wiki.php?edit=Library`

Clicking the "edit" link will bring up the same page again, but instead of converting the wiki page text to HTML for display, it is enclosed as plain text in a big textarea field in an HTML form, with a "save" button underneath. The reader can now edit the text and submit the new version, which will immediately replace the old version on the website. Any user can edit any page on a wiki website at any time.

Clicking the "save" button submits the form data to the wiki script, which stores the new text as a new version of the same wiki page, using a revision control system. Earlier versions and the differences between them are availble for review, using links from the normal wiki page display. A record is also added to a log of recent changes, and this list can be displayed by the wiki script, also using a link from the normal display of any wiki page.

Wiki pages are written in plain text, no different from sending ordinary e-mail. There is no need for the user to learn a cryptic code language like HTML. A blank line separates paragraphs. When a new page is saved, the wiki script translates it to HTML for presentation. In this process, any URLs are made clickable, and any URLs that end in .gif, .jpg or .png, indicating an image, are converted to an inline presentation of that image. A few other notations and shorthands are substituted as well, and the website maintainer can easily modify the wiki script to do the transformations of her choice. The idea is that the user doesn't need to learn all of these transformations. Editing plain text is useful enough.

The most important transformation, however, is the wiki page link syntax. This is the syntax used for creating hypertext links to other wiki pages. Typically, a word or phrase placed in [[double brackets]] is made into a link to a wiki page having that title. The exact syntax can vary with the implementation. Some wikis use WordsWrittenTogether as page titles, some use [single brackets].

Creating wiki page links is also the method to create new pages. If the word or phrase enclosed in brackets is not a title of an existing wiki page, a link is created that leads directly to the edit form for a new page having that title.

Most wikis also feature a search function, that performs a full text search among the pages of the wiki website. Several other functions can also be found. Since the software comes with open source, it is also common that the website maintainer has enhanced its functions and design.

4 Social and Legal Aspects

Most people, when they first learn about the wiki concept, assume that a website that can be edited by anybody would soon be rendered useless by destructive input. It sounds like offering free spray cans next to a grey concrete wall. The only likely outcome would be ugly graffiti and simple tagging, and any artistic efforts would not be long lived.

Still, it seems to work very well. I think the main reason is that the pages of a wiki are kept under version control, and anybody can restore a bad or ugly contribution to a previous version. From the list of Recent Changes, it easy for a regular contributor to keep an eye on new contributions, to view the difference between the new version and the previous one, to make corrections or additions to the page, or to restore the page to a previous version.

Another common fear is that different opinions on a topic would lead to editing wars. While this can be true to some extent, such wars consume a lot of energy, and most wiki contributors learn to take a neutral point of view as a way to avoid conflicts. Instead of writing that "apples taste good", they write "many people consider apples to taste good". A successful contribution is one that is left alone by others, one that expresses a consensus among the user group. For this, it must not only be neutral and objective, but also complete.

Within the Wikipedia project, the neutral point of view, or NPOV, as it is most often referred to, has been adopted as an explicit and official policy. Read more about this on:
`http://www.wikipedia.org/wiki/Wikipedia%3ANeutral_point_of_view`

In fact, a more real threat to a wiki website is that nobody wants to edit anything. The joy and usefulness of making contributions to a collaborative project needs to see the contribtions from others. It seems that an active core of at least five regular contributors are needed to keep a wiki alive. The first or first few individuals must be very determined in getting this process started.

Perhaps a more intriguing issue is that of copyright. After several people have contributed to an article, can anybody claim copyright to it anymore? And can anybody claim copyright to the collection of articles? As far as I know, this has not yet been tried in court, and copyright law clearly wasn't designed with this sort of composite works in mind. Traditional newspapers and encyclopedia are similar collections of contributions from many individuals, but the traditional employment relationship between the publisher and the authors of individual articles is not present in an open wiki. For intranet wikis, such as in a software develpment project, the collective outcome of the editing process is typically owned by the employer. But for volunteer based open source projects or publicly open wikis, the individual contributor has a stronger, more independent role.

Another side of the copyright issue is the use of copyrighted material in the wiki. If a contributor adds contents to the wiki that is covered by someone else's copyright, who is responsible for the possibly resulting damage? Swedish legislation, and I think that of most other countries as well, make a difference between edited newspapers, where the publisher is responsible, and bulletin board systems or web hosting services of an Internet service provider, where each user is responsible. The issue of responsibility also pertains to slander and hate speech. As far as I know, the applicability of existing laws to wiki websites has not yet been tried in court.

On these legal issues, Wikipedia has taken a preventive stance. Each contributor is granted copyright to their own contributions, but they are also informed that their pressing the "save" button constitutes an agreement to make the contents available under the GNU Free Documentation License (FDL). This license text has been written by the Free Software Foundation (FSF) for the purpose of granting to technical documentation the same freedom as the GNU General Public License (GPL) grants to software. In essence, that means that anybody is free to copy the text and use it for other purposes, provided they grant the next user access to the editable text (the source code). The easiest way to grant that access is to mention the Wikipedia URL where the contents was found.

From the above, it might seem that wikis are plagued with problems and risks. This is not at all the case. Everyday life on a wiki is full with the joy of learning and teaching, of seeing a collection of articles grow, all neatly hyperlinked without any broken links (links are created by putting words in brackets – they either lead to existing articles or articles that might come into existence) or uncontrollably complex hierarchical structures (all wiki articles exist side by side in a flat namespace). The user that has a different opinion about a topic can edit the page right away and see his contribution immediately published. Sometimes two different views on a topic will clash, and a minor editing war will ensue, similar to a discussion on a mailing list or a weblog. The difference between the mailing list discussion and the wiki editing war, however, is that the wiki leaves a more permanent record, the achieved consensus about the topic (or a documentation of the disagreement) that can easily be referenced from other articles. In this way, the wiki accumulates the experience of the user community, rather than just providing a forum for discussion.

Within a wiki, different users will take on different roles. Many will only read, perhaps only using the wiki as an occasional reference. Some will stick to their favorite set of topics. Some will frequently return to the list of recent changes, checking up on the contributions by others. While some might add large volumes of text, others can edit these tex~~ts to~~ add brackets and other forms of markup.

To measure the growth and size of wikis, various metrics have been developed. The most basic metrics are the number of pages and the number of contributions per time unit. Each contribution is a user pressing the "save" button, resulting in a line being added to the recent changes log, so the number of contributions can easily be counted. No metric has yet been devised for measuring the size of each contribution, to make a difference between a minor spelling correction and a large text submission. The most obvious metric for measuring the size of the entire collection would be to count the number of wiki pages. This simple page count, however has some drawbacks. It makes no difference between long articles and short ones, or even simple redirects such as from Rhodesia to Zimbabwe. One possible solution would be to count only those articles that are longer than, say, 200 characters or 15 words. However, the metric that actually gained momentum during Wikipedia's first year (2001) was instead to make a search for a comma and see how many pages were returned. This "comma count", as it has come to be known, is perhaps not the best possible metric. But as long as all compared wiki websites use the same, it works well for comparison. For most large wikis, it seems that the comma count is approximately 80% of the total number of articles. Wikipedia is currently developing these metrics further. In addition to these metrics, most website metrics (page views, unique visitors, etc.) would apply to wiki websites as well as to any other site.

By July 2002, the biggest known wiki websites are

1. Wikipedia, English version, 30,000 articles
2. C2.com, 18,000 pages
3. Enciclopedia Libre, University of Sevilla, 9,000 articles
4. Twiki (`www.twiki.org`), 8,000 articles
5. Susning.nu, 7,000 articles
6. NoSmoke (`http://no-smok.net/nsmk/NoSmoke`), 4,400 pages
7. German Wikipedia, 3,700 articles
8. Foxpro Wiki (`http://fox.wikis.com/`), 3,700 pages
9. Polish Wikipedia, 3,600 articles
10. Esperanto Wikipedia, 3,100 articles

5 Susning.nu

After learning about wiki and especially Wikipedia in the spring 2001, I understood that this technology could solve a number problems for me. I needed a better way to maintain the websites of a couple of volunteer based community projects that I have started over the years, and I also wanted to explore some ideas about timelines and web-based geographic information systems.

On August 31, I downloaded the UseModWiki software and installed it on my own site, the same version that was used for by Wikipedia at the time. I improved the Swedish translation of the software's text prompts, and submitted the same for use in the Swedish Wikipedia. Already on September 10, my wiki had 200 pages, all written by myself. October 1, 2001 marks the official opening on the website, which by then had moved to its current URL, `http://susning.nu/`

My early announcements on a couple of Swedish mailing lists attracted the interest of among others two Swedish pioneers in new technologies for libraries, Sigfrid Lundberg and Yngve Johnsson. Yngve was involved in a working group for a Swedish library web portal, and started to use my wiki site as his prototype system. Both contributed many articles on topics pertaining to the needs of libraries and librarians.

On October 4, susning.nu had 1000 pages, and the next 1000 pages were written in the next 20 days. The site already had some new functions in addition to the original UseModWiki software. The first of these are the search links in the navigation menu. Since every page has a title that describes its contents, it was very easy to generate a URL, using CGI URL syntax, that leads to a search for the same word or phrase on Google and some other search engines.

By the end of October, major search engines like Google and Alltheweb had indexed the contents of susning.nu, and started to deliver more visitors to the website. Google seems to like wiki websites, because there are many links to each page. Perhaps it helps that each page contains a search link back to Google. By July 2002, Google's cached version of susning.nu pages are seldom more than a few days old.

Among the automatic text substitutions that are part of most available wiki implementations is a recognition of the letters ISBN followed by ten digits. Where this pattern appears in a submitted text, a couple of links to online bookstores are inserted in the resulting HTML web page. For my wiki website, I changed this code to insert links to Swedish bookstores and for ISBNs starting with digit 3, also to the German Amazon.de. The combinations are endless.

When the susning.nu script finds a pair of parentheses containing the name of a foreign language followed by a colon and a word or phrase, it assumes this is a translation of the page title into the specified language. This is used for generating a list of direct and search links for the foreign phrase to a number of useful websites specific to that language, including the national language departments of Wikipedia, the Open Directory Project, and Google. For example, the page `http://susning.nu/Ryssland` contains (engelska: Russia), resulting in a link to `http://www.wikipedia.com/wiki/Russia`, where more information can be found on the same topic but in the English language. These translation links were introduced on November 25. Later, a similar system has been introduced in Wikipedia as well, known as language links. The concept is similar to Interwiki links, that already were in use by parts of the wiki community.

Mapblast.com is a website that kindly generates geographic maps as GIF images from geographic coordinates embedded in the URL. The exact economics behind this escape me, but they seem to provide this service for free for non-profit purposes. When the susning.nu server script finds the word "map" followed by three colon separated numbers in a wiki page, this results in an inline image of a map from Mapblast. The three numbers are the geographic latitude, longitude, and scale. In Mapblast's CGI URL syntax, this corresponds exactly to the CT attribute. In this way, maps can easily be included in the wiki pages in susning.nu, and the wiki becomes a gazetteer. By October 6, all Swedish cities and municipalities had been entered with a name, map, and link to the municipal website. Most countries and many cities of the world have been entered later.

The map function of susning.nu is primarily a way to present visible maps. But it is also in fact a way to encourage users to embed geographic metadata into the wiki pages. The visible map that is produced, provides instant feedback that the entered metadata are correct. But how can these metadata be explored? In January 2002, I implemented a geographic search function that looks for "map" patterns in all wiki pages and computes their geographic distance from a given starting point. It presents a list of the pages found, ordered by ascending distance from the closest to the farthest. Since this search function needs the geographic coordinates of a starting point, it was natural to use each map coordinate for this. Thus, the code that inserts the inline map image now also inserts a hypertext link to the search function, where the coordinates are submitted as the starting point. The link text reads "find other places nearby".

The geographic search function does not only use the distance between the geographic coordinates that represent the center point of each map, but also the map scale, since this is information about the size of the geographic object. Bigger objects are ranked higher than smaller ones in the hit list, so a search from Dresden will present a hit list starting with Europe and Germany, but a search starting from Europe will not have Dresden anywhere near the top. The exact ad hoc algorithm for this search hit ranking is a well kept secret.

Also in January 2002, a new set of pattern matching regular expressions were introduced, that recognize dates written in normal prose, including patterns like "July 14, 1789", "July 1789", "year 1789", "born 1789", "died 1789", "1780s" and "18^{th} century". When a date like that is found in a wiki page, it is made into a hypertext link to a date search function. This is a function that searches all pages for similar date patterns, computes their Julian day number, and presents a timeline of adjacent hits. The user doesn't have to learn any new syntax, but any mentioning of a date in one of these formats is automatically recognized. No users manual is provided, but the curious user is expected to click on the hyperlinked dates and discover how the date search function works.

The geographic and date search functions have turned out to work very well and be really useful. Any mentioning of a date or a geographic coordinate in a wiki page is metadata provided as a basis for smart searches. The usefulness of a metadata based or smart search function depends entirely on the quality of the metadata. The need to provide the general public with good incentives for entering quality metadata has in my mind been underestimated in some other systems, including the early hopes that Dublin Core metadata would gain widespread use on the World Wide Web. The map coordinates that susning.nu's geographic search function is based on, are entered for the purpose of drawing a map, not for the search function alone. The dates used by the date search are immediately useful in the text where they appear, not for the search function alone.

Before the end of 2001, susning.nu had reached 4000 pages. The full text search function from the original UseModWiki software as well as my initial implementations of the geographic and date search functions were based on a linear search through plain text files, one for each wiki page. Even though modern computers are amazingly fast, opening and reading 4000 files for every search was becoming annoyingly slow. The implementation has since been enhanced so all pattern recognition is made when a new version of a page is submitted, and the structured data are inserted into a MySQL database. After this, using the search functions is as fast as viewing any page. MySQL is also used for the full text search.

In March, 2002, susning.nu reached 6000 pages of which 4198 contained at least one comma, qualifying for the "comma count" defined by the Wikipedia project. By mid July, the site had 11000 pages of which 7800 pass the "comma count". It is believed to be the 5th largest Wiki website in the world. Another thousand pages are created every 20 to 50 days, and this has been the case since the opening of the site. Every month, the site receives contributions from between 50 and 100 registered users and a number of anonymous ones.

6 Future Directions

In this final section, I will touch briefly on some future developments that might come out of the wiki and related movements.

In parallel to wiki, another big contemporary movement is that of web logs, also known as blogs. A blog is more like a discussion forum, where each user can add a comment to a previous article, but a user cannot edit another user's text. Perhaps the most wellknown weblog in the open source software community is Slashdot.

Both wiki and weblogs represent a new trend in advanced, yet standardized server-side software. Only limited technical knowledge and a very small budget is needed to set up a wiki or weblog, especially compared to commercial web content management systems (CMS) that are equally powerful. Wiki and weblogs are disruptive technologies and the young CMS software industry better watch out.

It is an addictive convenience to have a simple search function available in the navigation menu of each page. In the case of a combined website that might feature both a wiki, a weblog and static contents, the same search function should cover all contents available on the web.

Advanced pattern recognition for written text, similar to the date search function that I implemented for susning.nu, could be applied to any text corpus, including webpages harvested by global search engines like Google or Altavista.

The typical front page of a blog website is a list of the most recent top level articles, with links to pages containing the comments. This is similar to the recent changes list of a wiki website, and a central place that frequent visitors return to. A special data format known as RSS can be used to harvest a summary of the most recent headlines from a blog site, and several wikis (including susning.nu) provide a similar feature. Such summaries from different sources can be compiled into a personal daily digest, using advanced RSS client software. Other websites could also benefit from providing such RSS feeds, including press releases. This could be the first step towards a more semantic web.

Both wikis and weblogs allow users to contribute text using their web browsers. This text can be in HTML, if the users care to learn HTML, but many sites allow plain text entry using a simplified markup that the server-side script can translate into a presentable HTML web page. This plain text format has no name and is not standardized. In time, the formats used by the largest websites and most commonly used free software implementations might become formally standardized. Such a format could use a more general framework for writing hypertext link shorthands, inspired by the KDE web browser Konqueror, where the "gg:" prefix is used as a shorthand for a Google search, just to mention one example.

Alternatively, it could be hoped that web browsers would come to support WYSIWYG editing of HTML forms in some richer text format than just plain text, and in general behave more like advanced text editors.