# Great Changes in Wikipedia

*by Lars Aronsson (user:LA2) lars@aronsson.se, October 2008*

## *Wikipedia is great. But will it remain so?*

Wikipedia, the free encyclopedia, is now more than seven years old. It has gone through several phases of development and increasing maturity, in technology, content and social relations between users. New features and traditions have come into use, and have become the established way of doing things. Radical changes appear more seldom, and newcomers are sometimes told that the current way of doing things is "how it always was".

This conservatism is natural in an encyclopedia, trying to accumulate, contain and present the sum of established human knowledge to a younger, learning generation of readers. But an encyclopedia is also a revolutionary tool, providing the power of knowledge to the young and powerless.

Wikipedia was perhaps the first major encyclopedia project that was "born digital", not an effort to adopt an old printed encyclopedia to modern information technology, but originally created for the World Wide Web. It doesn't use the web merely for reading, but also for writing, with a free license to make the content reusable. This is a radical step forward in the already radical history of encyclopedias.

But as Wikipedia grows, becoming more complete and improving in quality, it becomes harder to make the encyclopedia itself subject to radical changes. It is often claimed that anybody can edit Wikipedia, and this is somewhat true for individual articles, but the overall structure is increasingly hard to change.

## *Size*

Wikipedia is like a word processing document that we edit together. If we find a spelling error, we can search-and-replace it. On my first home computer, I could write documents that were 30 kilobytes because that's how big the memory was.

The English Wikipedia now has 2.5 million articles and there are 23 languages with more than 100,000 articles. The 23rd biggest is the Slovak Wikipedia. The average article is roughly 2 kilobytes. This means the English Wikipedia is a corpus of more than 5 gigabytes and the Slovak is 200 megabytes.

The whole text of Wikipedia (current versions only) in any language can easily fit on the disk of any personal computer, but there is no practical way to make a quick search-and-replace over the whole corpus of text. Instead we retrieve a single article at a time, some 2 kilobytes on average, in an HTTP call, make modifications and submit it back to the Wikipedia web server. Current robot guidelines allow one edit per ten seconds. If we don't need to pause or think, but can work 24 hours per day, going through the 100,000 articles of the Slovak Wikipedia would take 12 days and the 2.5 million articles of the English Wikipedia would take ten months.

## *Big and small changes*

Some changes in Wikipedia can be small and still very important. For example, creating a new guideline to say that politics must never be mentioned or that all articles about plants must have a photo, would have a huge effect. Creating the guideline is just a single edit. The discussion of whether such a guideline is useful could probably last a year. But after this, the actual implementation, removing all texts about politics or finding photographs of every plant, could take for ever.

Even if consensus is reached for such policy changes, if the implementation is delayed, this might cause the community to reconsider and the policy might be revoked. What this means is that community consensus is actually less important than technical feasibility. Technical reality has a veto against community consensus.

There are three forces at work here: 1) New visionary ideas on how to improve or change Wikipedia, 2) the consensus-building opinion of the other contributors, and 3) the technical feasibility of implementing the changes. This paper focuses on the third force, but also considers the second.

## *Examples from the Swedish Wikipedia*

My background in Wikipedia (user:LA2) goes back to 2001. I also operated by own wiki website "susning.nu" in 2001-2004, which has since remained closed. I have been active in many languages, but never reached a high edit count until this year. In 2008 I have been on the board of Wikimedia Sverige, the newly established Swedish chapter of the Wikimedia Foundation, and also spent more time on Wikipedia.

## *Substub hunting*

For the Swedish Wikipedia, 2008 started out with an essay on quality improvements by Lennart Guldbrandsson (user:Hannibal), who is also the chairman of Wikimedia Sverige. In its early years, Wikipedia has fought to create a large number articles, and the Swedish Wikipedia was very successful in this. It has a large number of articles, but many of them are very short and not very good.

I am among those who worry that many short articles will give readers a bad impression of Wikipedia's quality, and also set a bad examples for new contributors who might think that short articles are okay.

Some contributors to the Swedish Wikipedia do not consider this to be a problem, either because they are optimistic that short stubs can encourage users to improve them and in time all will become better, or because they have not fully realized how much shorter the average article is in the Swedish Wikipeda, compared to other languages.

Policies on the German Wikipedia call for very short stubs to be removed. That would be a quick solution to the problem. But such a policy is not in effect on the Swedish Wikipedia, and suggestions are met with strong opinion.

I'm a programmer with a UNIX background. I like to analyze the database dumps from Wikipedia, because this allow me to analyze the entire content in ways that the website or API doesn't provide. Even if changes cannot be made as fast, at least I can get a better understanding of what needs to be done.

Some years ago, I wrote a little Perl script "extraktor.pl" that extracts data from template parameters in Wikipedia articles, http://meta.wikimedia.org/wiki/User:LA2/Extraktor

A slight modification of this script gives me useful statistics on the type and length of articles.

The median length of articles on the Swedish Wikipedia increased slightly from 972 bytes in February to 1058 bytes in June 2008. Neighboring languges Norwegian, Danish, Polish, Finnish and Czech all have larger values, ranging from 1381 to 2195 bytes. The German and Russian Wikipedia have 2740 and 3572 bytes, respectively.

But more telling than the median length is the lowest decile, i.e. the length in bytes of the 10% shortest articles. For Swedish, this was 298 bytes in February and increased to 349 bytes in June 2008. However, for Danish, Norwegian, Polish, Finnish, and Czech this value ranged from 491 to 917 bytes. And for German and Russian, it was 1081 and 1305, respectively.

It seems that more mature Wikipedias have a smaller fraction of very short articles. The many very short articles in the Swedish Wikipedia is typical of much smaller language versions of Wikipedia, such as Estonian or Arabic. But even in the Estonian and Arabic Wikipedia, the 10% shortest articles reach beyond 400 bytes.

Many people associate very short articles with robots creating new articles based on very simple statistcs, such as the population of small villages. However, this has not been the case in the Swedish Wikipedia. Some substubs are about villages or geographic districts and some have been created very quickly in large numbers during Wikipedia's early years. But they have been created manually, not by robots. In fact, robot-created articles, such as those created in 2002 in the English Wikipedia by the Ram-Man robot, are generally of higher quality than the Swedish Wikipedia substubs.

## *Finnish villages*

By looking at my statistics, I could identify some important groups of substubs. One such group was villages in Finland. The articles typically only stated that "Klobbskat is a harbour village in Korsholm municipality". My suggestion to erase such meaningless stubs was met with wild protests. No information was allowed to be lost. Instead, I grouped this listing by municipality and inserted a list of village names in the article for each municipality. I then replaced the substubs with redirects to their municipality.

This work was mostly manual work, based on lists from my statistic analysis. The removal of 1500 such substubs lasted for the month of April 2008. This speed equalled the creation of new articles, and the article count didn't change for the whole month.

When I had taken care of all Finnish villages having stubs shorter than 160 bytes, it was already May.

## *Swedish week*

For a long time, the Russian Wikipedia had been the 11th biggest while Swedish was number 10. The top 10 languages are listed around the Wikipedia logotype on the www.wikipedia.org front page, and the Russians were now eager to reach this position, which they finally did on May 19. The last ten days was celebrated with a "Swedish week" campaign, where they created new articles pertaining to Sweden.

## *Categorizing substubs*

While the villages of Finland were easily handled by merging substubs into lists of village names in the articles of each municipality, I didn't care to do the same for subtubs on other topics.

The Swedish Wikipedia has various WikiProjects and one of them addresses substsubs. In March 2008, the project page said there were a total of 233 substubs in the Swedish Wikipedia. My conclusion from the statistic analysis was that perhaps there are 23.000 substubs. But this project talked about articles that had been put in the category:substubs by inserting the substub template.

Most languages of Wikipedia have at some point organized stubs in categories using templates, an activity called "stub sorting". Some languages have abandoned this practice, since they deemed in counterproductive. But it still lives on in the Swedish Wikipedia, including a special category for substubs.

The intention behind the Swedish substub template was that a new category should be created for each month. If a substub was categorized as such in September, and had not improved by November, perhaps it could just be erased. However, the template implementation was broken, so all substubs belonged to the current month. And nobody followed the categorized substubs.

What I did was to fix the broken template implementation, and I changed the interval from months to weeks. Then I started on April 24 to add the substub template to those very short stubs that weren't villages in Finland. While I started doing this, I was categorizing some 200 substubs every week. I could soon prove that there were far more than the originally claimed 233 substsubs in the Swedish Wikipedia.

In June I added 500 substubs in one week, and this was too much for some project participants, who asked me to stop. By then, 1673 substubs were waiting in 8 categories, one for each of the previous weeks. The number was only decreasing very slowly, by 50 per week on average. Even if the markup of new substubs stopped, the backlog would take almost a year at the current pace.

Still, a backlog might not be a big problem. The idea was that substubs that hadn't been addressed (extended or merged into real articles), could be removed after some time. No final date was decided, but some project participants did occasionally remove substubs that they didn't care to improve.

However, when one project participant in October erased all 31 remaining substubs from a category that was created 20 weeks earlier, some users objected in a loud voice on the Swedish Wikipedia village pump. They had never heard of this massive removal of articles. There was no peace before the 31 substubs were reinstalled.

This is a problem of the wiki model. Anybody can edit, and anybody can enter the meeting at the last minute and have a voice about things that were discussed and decided months earlier. Consensus back then, doesn't matter here and now. And wiki is all here and now.

The conclusion must be that it is useless to timestamp the categorization of substubs and other defects. In marking up an article that needs to be fixed, you can allow some time for this to happen, but after that time has passed, anybody can still disagree that this time was enough. And their opinion, no matter how unfounded, can be allowed to break the consensus you thought you had.

If allowing time for improvement doesn't work, immediate action does. While I was merging stubs about villages in Finland, I did have plenty of hard work, but I didn't have conflicts of this kind.

## *Years in sport and film*

Another category with many short articles were year chronicles. There is a pattern in every language of Wikipedia to have an article named "1973" for events of that year, people who died and were born, and important (Nobel) prizes that were awarded. For some languages, there are also specializations such as 1973 in sports and 1973 in film.

The Swedish Wikipedia had 20 such topics for each year, which is a bit much. Many of these topics had very little content, meaning that there were gaps in the chronology and many articles were short stubs that only mentioned a few events for a year. In order to learn something from an article about "1973 in economics" that article needs to list a dozen or more events and they need to be the important ones, not just some random events that somebody happened to list.

In June I started to try to understand how to address this. There was "1983 in IT" (information technology) that I could easily merge into "1983 in technology", but this just covered a few decades. All the years "in technology" before 1950 were very short, so I merged them into years "in science". The English Wikipedia doesn't have a separate topic for years in technology, anyway, so those didn't have any interwiki links. The longer I worked, the more I realized that I needed to address all topics and all years.

I late July I listed all 19 topics (now that IT years were gone), which years they spanned, whether they had interwiki links, and how many stubs they contained. I used CatScan to find stubs and set the limit to 512 bytes. In June these 19 topics contained 1016 stubs. By the end of July I had already broght this down to 845. I made it a habit to list the stub count every Monday morning. Each Monday, I calculated the previous week's reduction and estimated the number of weeks remaining if I could keep pace. At the end of September there were 400 stubs remaining. The years "in education" and the years "in economics" both covered the 20th century but contained 85 stubs each, which all disappeared when these topics were merged into the regular years.

Some topics were extended and improved, rather than being merged. This included finding events and births and deaths for each year. It also meant filling gaps in the chronology. In August I started a similar log for the number of remaining gaps.

The years "in science" used to start from 1600. In July this was changed to 1700, leaving only 33 stubs and 44 missing years. By August the stubs were all improved beyond the 512 byte limit and by September all gaps were filled. As soon as I started to extend the years in science, a few other users came to help. From being a stub improvement project, this now turned into a science history project.

Progress has been slower for other topics. In October 2008, there are still 63 stubs for years "in art" and 59 for years "in radio". Other topics have fewer stubs, but the total for all years are 396 stubs. While this sounds bad, it is a great improvement over the 1016 stubs that existed in June. The number of topics has shrunk from 20 to 17. Much work has gone into extending and improving articles for years.

## *Gender categorization*

Categories were introduced in 2004 when Wikipedia was already three years old. They are used in almost the same way in all languages, with one important exception. The German Wikipedia community had the idea that categories should be atomic and orthogonal, assuming that cross-category search would soon become available. Thus, instead of the usual "category:Canadian actor" they have one Category:Canadian and another category:Actor. To find the Canadian actors, you need to find articles that use both categories. Since this kind of search is not provided by the MediaWiki software, an external tool is needed. This is provided by the CatScan application on the German toolserver.

We can imagine what would happen if MediaWiki suddenly did provide cross-category search. The German pattern has a certain elegance, maybe some more languages would want to switch to this. Or on the other hand, given the current state, the Germans might want to abandon their failed system and adopt the internationally accepted system of combined categories for nationality and profession. But how much work would that be? Even if there would be consensus, can it be done?

Time is a factor here. If there is consensus and the work is started, can the transformation be completed fast enough, before the consensus swings back? If the speed drops, leaving the work half done, it might be easier to return to the previous well-known state than to go on.

Most languages do apply such orthogonal categories for people's birth and death year. The "category:1973 births" is always kept separate from a person's nationality and profession. There is no category for Australian tennis players born in 1973, not even for Australians born in 1973.

One set of categories that the German Wikipedia is alone in applying, are those for men and women. Every article describing a person is not only categorized by nationality, occupation, and birth and death year, but also by gender.

The English Wikipedia in some cases have gender as categories. For example, the category:Sportspeople has a subcategory for Sportswomen, with further subcategories for Female divers and Women cricketers, even Canadian women's ice hockey players. But this is not a system that includes all biographic articles, such as the categories for births and deaths. That is different in the German Wikipedia. If you add all 34,440 women and 200,886 men and the 4 intersexual people, you should have every biographic article in the German Wikipedia.

As the reader can see, one effect of having these categories is that Wikipedia's gender bias becomes painfully obvious. There are six male biographies for every female in the German Wikipedia. Nobody knows what the ratio is for other languages, because nobody has counted.

Having these categories, you can use CatScan to find out how many of all actors or physicians or Nobel prize winners are male or female.

Another less obvious advantage is that the categories do document what gender each person has. Sometimes I read an article that "Kim Foo is a singer". If the forename is ambiguous (such as Kim) and the article text never refers to the person as "he" or "she", the reader can't know if this is a man or a women. The gender is obvious from some titles (actor, actress) but not for others (singer).

At the end of August 2008, I decided to implement categories for men and women in the Swedish Wikipedia. I had long considered this, and thought that it would be possible. But it was an experiment. I copied the system from the German Wikipedia, including the third category for intersexual people.

In the German Wikipedia, 28 percent of all articles are biographies. If the Swedish Wikipedia has the same ratio, and should it not, then I should expect 80,000 biographies.

To begin with I had extracted a list of 50,000 article titles from the available database dump that contained either a birth or death category. I sorted this list alphabetically and started with all named Maria and Anna. Using the pywikipedia robot framework, I quickly looked at each article and added the category:Women. Initially I made some mistakes. A Swedish church parish named Maria and the artist Alice Cooper were both categorized as women. Soon I made some changes to the pywikipedia script to avoid such mistakes.

Pretty soon, protests were heard. Discriminating people based on gender is not allowed, and even documenting the gender was enough to stir emotions. I was able to meet these protests with good arguments, saying that the gender is already contained in biographic articles titled Maria or Peter, just not available to searching. And by making it available to search, the existing gender bias can be exposed.

Another protest concerned the intersexual category. There was a community decision to remove it. Intersexual people are instead categorized as both man and women. This is also how the German Wikipedia does in some cases. There was a discussion of which definition of gender to use, and the conclusion was to use the legal definition rather than the person's own aspiration.

Again, I made it a weekly habit to document my progress. In the second week, three more people got copies of my modified pywikipedia script and started to help me. After four weeks, we had categorized 35,000 biographies. After six weeks we reached 50,000 biographies. By then, there were already 3 men for every woman. But since I had started out with women, the ratio of men is still increasing. In the seventh week 7100 men were added but only 479 women.

In the first two months, far more than half of all existing biographies have been categorized. Some weeks remain. But already, the gender categorization has been very successful. After it was envisioned, the initial protests were taken care of and the technical solution proved to be fast enough to get the work completed in reasonable time. More people started to help with the project. Others have taken up the habit to add these categories to new articles.

## *Conclusions*

Anybody can edit Wikipedia, at least individual articles. It turns out that greater changes are also possible, but they don't come easy. The size of the text corpus and the number of articles make great changes take weeks rather than hours. During this time, community consensus can change, and this makes swiftness important.

If you try to implement great changes in Wikipedia, beyond trivial edits of a few dozen articles, you are in a similar dilemma as described by military theorist Carl von Clausewitz: A general fighting a war not only has to handle an enemy army, but also the swinging opinion of his own government. A sudden order to withdraw can be as disasterous as an enemy victory on the battlefield.

In this case, your government is the consensus among the user community. If you act swiftly and in a predictable way, reporting your progress on a weekly basis, chances increase that the community will keep faith in your actions.

One of the best technical tools for great changes in Wikipedia is the availability of fresh database dumps at regular intervals. But this has been failing during the summer and fall of 2008. No dumps were produced during July, August and September.

## *The next step*

I already see the end of gender categorization. There are a few topics of year articles that should be extended and completed. There are still many short stubs that need to be taken care of, and the next database dump will indicate just how many. However, categorizing them as substubs is not a viable option. Instead, the approach taken with villages in Finland should be followed. It is necessary to group the stubs by mother articles, into which they can be merged.

I think my next major project in the Swedish Wikipedia will be to add geographic coordinates to articles about places. Not only will this provide the ability to locate them on a map, it will also indicate which articles describe places. That would be in line with the gender categorization, which indicate which articles are biographies.

## *References*


[[sv:Kategori:Geografistubbar-Finland]]

[[sv:Wikipediadiskussion:Projekt substubbar#Årsartiklar]]
[[sv:Wikipediadiskussion:Projekt infogningar#Årtalsartiklar]]
[[sv:Wikipedia:Årskrönikor#Fördelning på ämnen]]
[[sv:Wikipediadiskussion:Årskrönikor#Stubbstatistik]]

[[sv:Mall:Substub]]
[[sv:Wikipedia:Projekt substubbar#Artiklar i Kategori:Substubbar]]
[[sv:Wikipediadiskussion:Projekt substubbar#Framsteg]]

[[de:Kategorie Diskussion:Mann#Statistik]]
[[sv:Kategoridiskussion:Personer_efter_kön#Är vi klara snart?]]