

Keep it Simple: Editorial control at simple.wikipedia.org

Matthijs den Besten (University of Oxford)

&

Jean-Michel Dalle (Université Pierre & Marie Curie)

Background

The WikiWiki collaborative editing system originally created by Ward Cunningham in 1994 [1] has become a very popular and widely used tool in recent years, most notably for Wikipedia, the free online encyclopaedia, but also within various organizations for project communication and documentation. The success of Wikipedia, its rapid growth and visibility on the Internet, provides evidence that WikiWiki works well. Yet, it is also this particular application of wiki technology to the encyclopaedia that draws most criticism, for instance from star Computer Scientist David Parnas: “Wikipedia provides a fast and flexible way for anyone to create and edit encyclopedia articles without the delay and intervention of a formal editor or review process,” he acknowledges. “But will this process actually yield a reliable, authoritative reference encompassing the entire range of human knowledge?” he asks [2].

People involved in Wikipedia generally seem to recognize that these concerns and several mechanisms are being tried out to alleviate them. It is these mechanisms for quality control that we will study here. More specifically, there are three broad types of modifications to the original WikiWiki system that have been added to Wikipedia so as to create a partial substitute to the formal edition and review process that is characteristic of traditional encyclopaedias:

- 1) The allocation of specific rights to a limited group of experienced users such as the right to block other users who are misbehaving and to the right to remove edit-possibilities to pages which are contentious;
- 2) The introduction of *bots* – “automatic processes that interact with Wikipedia as though they were a human editor” (<http://en.wikipedia.org/wiki/WP:B>) –, which carry out routine searches for, among others, page vandalism and broken links;

- 3) The application of labels, for instance to indicate that a page is of high quality (“featured article”) or that a page is considered for deletion (“articles for deletion”).

It is this in last modification that we are most interested for now. Do these page labels act as attractors for contributions or contributors? Do they generate more edits and discussions? And do the labels result in a higher quality of the pages they have marked? In a first step, we focus on the study of `simple.wikipedia.org`, “a free encyclopedia written in simple English for easy reading” which, with slightly more than ten thousand articles, is a relatively small offspring of the main Wikipedia; within this corpus, we look at the effect that the crucial label “*unsimple*” has on the readability of the articles that it marks.

This research is in a way a direct continuation of previous research that focused on what attracts developers to contribute to specific files in open source software systems [3]. Furthermore, the fact that editorial controls affect writing style has already been demonstrated by Emigh and Herring [4] in a comparison of Wikipedia with Everything2, while Les Gasser and colleagues have proposed a set of metrics for quality assessment that we happily adopt here due to its relevance to our topic [5]. For obvious reasons, for now, we specifically focus on readability metrics as measures for the simplicity of pages.

Preliminary Investigation

Our database contains revisions of 27 497 pages of which only 250 have been labelled “unsimple” at a point in their lifetime. These 250 pages however represent about one fifth of the size of the dump: this may be explained by the *larger average size of pages once labelled “unsimple”* and/or due to the significantly higher number of revisions on those pages (the “unsimple” pages have 736 characters and 37 revisions on average whereas the average over all pages is 159 characters and 6 revisions respectively).

The first question we wanted to investigate was whether we could establish a relation between readability measures and the appearance of the “*unsimple*” label on a page. For that purpose we applied the GNU style tool (version 1.02) to all revisions of all pages in our database and aggregated the results in the table 1 on the next page. For a variety of readability grades, this table compares the overall readability of the project with that of pages just before and after the tag “unsimple” appeared and disappeared.

Table 1 Effect of the *unsimple*-label on readability.

	Overall Median	Before Tag	p	After Tag	p	Improvement	p
Kincaid	8.39	11.30	0.05	8.65	0.79	2.52	0.02
ARI	9.25	12.94	0.05	9.79	0.68	3.01	0.03
Coleman	10.67	11.50	0.00	10.61	0.75	0.85	0.00
Flesch	68.74	58.69	0.01	69.02	0.92	-9.80	0.00
Fog	11.76	14.50	0.07	11.68	0.94	2.68	0.02
Lix	40.23	47.83	0.05	39.83	0.88	7.55	0.01
SMOG	9.67	11.36	0.00	9.83	0.56	1.44	0.00

More in particular, for seven readability grades listed in column one, column two gives the median score of all revisions of a page averaged over all pages. Next, the scores are given for the average state of pages just before the label or tag “unsimple” was applied and just after it was removed. The small values of p right next to the scores are the p -values of t-tests as an indication of the probability that two distributions are the same. Thus, the low p -values next to the column “Before Tag” indicate that, on the basis of a two-sided t-test, the readability of pages just before “unsimple” appears is significantly different from the overall readability in simple.wikipedia.org. In contrast, the high values next to “After Tag” indicate that pages can hardly be distinguished from regular pages anymore after the “unsimple” tag has been removed, while the low p -values on the right confirm that the “unsimple” tag yields significant improvements, at least for a one-sided paired t-test. The metrics shown – i.e. the Kincaid formula, the Automated Readability Index (ARI), the Coleman-Liau formula, the Flesh reading easy formula, the Fox Index, the Lix formula, and SMOG-Grading – are all well established readability grades with each their peculiarities. Hence it is all the more surprising that all point in the same direction. Especially neat in this context are the values of the Flesh reading easy formula. According to the manual page of “style”, “[t]he index is usually between 0 (hard) and 100 (easy), standard English documents averages approximately 60 to 70.” Not only are “unsimple” pages objectively significantly less readable than regular pages, and not only does the “unsimple” tag generate a significant improvement to a page in many cases, but, as if by magic, the community of simple.wikipedia.org manages to keep the readability of pages more or less on the simple side within the margins of standard English.¹

¹ NB: Values in Table 1 concern the 99 pages that were “simple” before *and* after.

Table 2 User activity by type.

	All pages	Unsimple pages	Edit while "unsimple"	Tag "unsimple"	Untag "unsimple"
Anonymous	46491 (26%)	2238 (37%)	745 (33%)	41 (16%)	37 (37%)
Registered	59235 (34%)	2073 (34%)	670 (29%)	98 (39%)	53 (52%)
Administrator	32387 (18%)	1248 (22%)	429 (19%)	111 (45%)	10 (10%)
Bot	38737 (22%)	504 (8%)	444 (19%)	0 (0%)	1 (1%)

Another question we investigated was whether the editors involved in tagging pages “unsimple” are any different from regular users of simple.wikipedia.org. Table 2 gives an indication of these differences and similarities. The table gives the rough counts of all revisions that can be attributed to anonymous visitors, registered users with an account, users with special “administrator rights” and software bots.² What is appears from this table is that the acts of adding and removing the “unsimple” label to and from pages is dominated by, but not limited to, administrators. Meanwhile, so far, *bots* are completely absent from the process, even though they play an important role in general. Is this the result of a conscious decision not to rely on automated processes for the potentially contentious nature of tagging? Or are the conscientious editors at simple.wikipedia.org simply not aware of the readability metrics that appear so powerful in predicting whether a page would be considered “unsimple” or not from Table 1?

Further research is needed to determine whether the efforts to keep it simple at simple.wikipedia.org are really as efficient as they appear to be at first sight and whether they could be improved. And afterwards similar questions can be asked for the broader issue of quality control in much large projects like en.wikipedia.org.

References

- [1] W. Cunningham and B. Leuf. *The Wiki Way. Quick Collaboration on the Web*. Addison-Wesley, 2001.
- [2] P. Denning, J. Horning, D. Parnas, and L. Weinstein. Wikipedia risks. *Communications of the ACM*, 12(48), December 2005.

² User groups were determined on the basis of a recent table user-group assignments. However, keep in mind that the assignment of groups to users need not be static.

- [3] M. den Besten, J.-M. Dalle, and F. Galia. Collaborative maintenance in large open-source projects. In E. Damiani, B. Fitzgerald, W. Scacchi, M. Scotto, and G. Succi, editors, *Open Source Systems*, volume 203, pages 233–244, Boston, 2006. IFIP International Federation for Information Processing, Springer.
- [4] W. Emigh and S. C. Herring. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences (HICSS-38)*., 2005.
- [5] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality*, 2005.