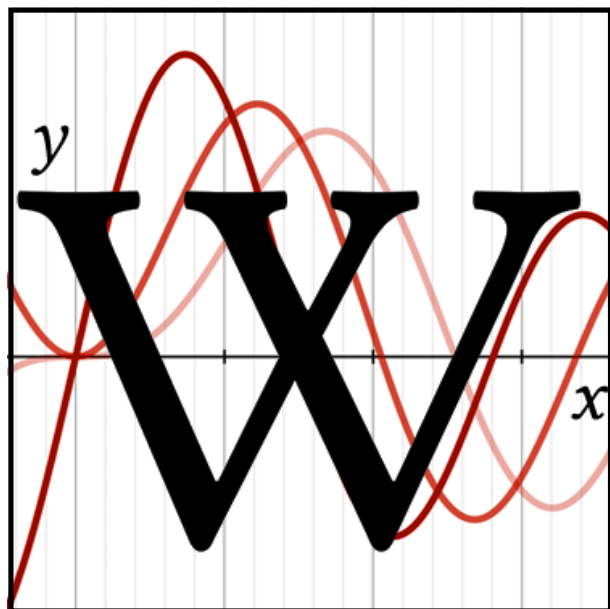


---

# Wikimedia Research Newsletter

VOLUME 1 (2011)

---



# Contents

About	1
Issue 1(1): July 2011	3
Issue 1(2): August 2011	12
Issue 1(3): September 2011	18
Issue 1(4): October 2011	24
Issue 1(5): November 2011	31
Issue 1(6): December 2011	35
Article Sources and Contributors	43
Image Sources, Licenses and Contributors	44
License	45


# About

The **Wikimedia Research Newsletter (WRN)** is a joint initiative of the Wikimedia Research Committee and the Signpost to cover research updates of relevance to the Wikimedia community. The newsletter is edited monthly and features both internal research at the Wikimedia Foundation and work conducted by external research teams. It is published as a section of the Signpost and as a stand-alone article on the **Wikimedia Research Index**.

## Facts and figures

The inaugural issue of the WRN was published on July 25, 2011, after two Signpost articles covering recent Wikimedia research. The six issues published in the first volume (July-December 2011) featured a total of 93 references and attracted altogether more than 17,000 pageviews (not including visits to the WMF blog edition).

## How to subscribe

You can subscribe to the newsletter via the following RSS feed from the Wikimedia Foundation's blog: 

The table of contents of each issue is cross-posted to [wiki-research-1](#)<sup>[1]</sup>

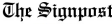
## How to contribute

This newsletter would not be possible without contributions from the research and Wikimedia community. We welcome submissions of new projects, papers and datasets to be featured in the newsletter. Work on the upcoming edition is coordinated on an Etherpad<sup>[2]</sup>, where you can suggest items to be covered, or sign up to write a review or summary for one of those that are already listed. Beyond that,

- If you want your **project** to be featured, please create a new project page using the form on the research project directory
- If you have published or you know of a recent **paper** that should be featured, please add an entry to the canonical directory of academic studies of Wikipedia
- If you have released **code** or **data** of relevance to research on Wikimedia projects, please contact us

For anything else (such as events, CFPs, research blog posts) please get in touch or make sure you post an announcement to [wiki-research-1](#)<sup>[1]</sup> (we are monitoring this list on a regular basis)

We are also looking for **contributors** (either occasional or regular) for the newsletter. If you have reviewed recent Wikipedia literature or would like to help writing the newsletter, please contact us.



25 July 2011

---

**RECENT RESEARCH**

**Talk page interactions; Wikipedia at the Open Knowledge Conference; Summer of Research**

By [Julia Abella](#), [Leticia Diaz Garcia](#), [Brenn Wolke](#), [David Mitchell](#), 30 July 2011 [Share this](#) [Print](#)

This is the first occasional review of recent published research on Wikipedia and other Wikimedia projects (previous issues: [June](#), [April](#), [Jan](#)). In addition to focus on covering research by academics outside Wikimedia, this issue features contributions funded by the Foundation itself. If you want your research to be featured in this monthly newsletter, you can tell us about your work by submitting it to the [Wikimedia Research Index](#).

**Editorial and conflict metrics**

A study covered in the previous edition of the research newsletter was extended and published by the authors on ArXiv. The authors report a new method for classifying how disputed a Wikipedia article is, to detect controversies and edit wars. At its core, the method is based on looking at sets of editors who have mutually received each other, and using their respective edit counts to define an overall metric of conflict. Even though this formula is not immediately intuitive, the authors describe using special diagrams called "heat maps" on the Conflict Index that spread across pages of editors. The authors use this procedure to select two samples of pages, of disputed and non-disputed topics, respectively, and analyse the time-series of revisions to those pages, which they find both time series are characterized by bursts of user activity. They claim there is a qualitative difference between the two, although their analysis appears to lack any form of statistical hypothesis testing. They apply a priority based model of editor activity that has been already proposed to explain human activity on the web, and find two distinctive patterns of activity that can help class "good" guys vs "bad" guys. [View on ArXiv](#)

**The anatomy of a Wikipedia talk page**

Several papers over the past months have focused on the structure and nature of social interaction on Wikipedia's discussion pages, both from quantitative and qualitative perspectives.

- **Wikipedia discussion relative to geography and history, but deep in philosophy, law, language and politics.**  
A study conducted by a team of researchers based in Italy and discussed and presented last week at SIGCHI '11 took into the perspective of the social interaction of participants in discussions on talk pages. The paper highlights a number of interesting issues in studying social network patterns in Wikipedia. Social ties in Wikipedia are explicit, unlike as there is no recommendation or explicit link between two Wikipedia users. A conversation between users allows retrieval of the typical social network. However, viewing such networks in Wikipedia is challenging for two main reasons: the lack of structure of talk pages which make conversations hard to control, and the dispersal of discussion threads, both within a page and over multiple pages (e.g. an article talk page can have a variable number of previous user talk pages). Despite these difficulties, the study analyses the properties of two types of social network: evidence article discussions (those on article talk pages and those that focus on an article but take place in media or user talk pages) and a semi-structured social network (i.e. the network derived by direct messages left by users on their talk pages). The three networks show interesting disparities in terms of the in and out degree of their nodes and in the proportion of working between their edges, suggesting that user and article-oriented communications are supported by substantially different networks.


The paper moves on to examine the degree *assortativity* of these networks - the tendency of nodes to connect with other nodes having a similar number of links. A striking difference emerges in the comparison with conversations in [Facebook](#), which are characterized by strong assortativity, and discussion networks in Wikipedia, which display a systematic *disassortativity*, an indicator of the specificity of social interactions in Wikipedia compared with other social media. As the authors summarize, "Wikipedians who reply to many users in article talk pages tend to interact mostly with users having connections, i.e. nodes and interconnected ones, while the Wikipedians who receive replies from many users tend to interact preferentially with each other". The study moves on to compare the depth and quantity of article-oriented networks of article-oriented networks and article-oriented networks of those discussions based on their depth and the number of initial replies among users participating in the same threads. The research characterizes the local frequency structure of discussions across different article categories and finds that although "Geography" and "History" attract together for almost half of all discussions in the English Wikipedia, they tend to involve female, whereas "Philosophy", "Law", "Language" and "Religion" are characterized by the opposite: female and receive the largest number of participants.

Two of the authors gave a presentation at last month's [European 2011 conference in Edinburgh](#): "[Co-authorship, 3.0: patterns of collaboration in Wikipedia](#)".- **Building consensus in talk pages: authority and alignment.**  
A group of researchers based at the [University of Washington](#) released an annotated corpus of discussions from Wikipedia talk pages including two years of article talk, alignment moves and authority users. In the authors' own words, "We authority claim to a consensus made by a discussion participant aimed at balancing their credibility in the discussion. An alignment move is a statement by a participant which identifies positions from all speakers or aligning with another participant or participants regarding a particular issue". Building consensus with the use of authority and alignment can help to shed light on consensus building strategies used by participants in Wikipedia discussions. The authors claim that the above information can be used to build tools to produce consolidated versions of online debates. The data spans 366 discussions that occurred in the last year between 2009 and 2010, involving a total of 1,616 editors. After presenting the corpus, the study presents an analysis comparing editor activity metrics with the propensity of adding one of the above social moves. The authors conclude on their findings: "our research is limited to the present, i.e. claim that the editor has made at least a move within the past 3 months, and report that this indicator of editor activity positively correlates with the propensity of alignment moves in discussion. Making an authority claim makes a user significantly more likely to be the target of an alignment move within the subsequent 10 turns compared to turns that do not contain any claim".
- **Recommending in the design of Wikipedia talk pages.**  
Researchers from the [National University of Ireland, Galway](#) presented work in progress from a project aimed at understanding [Wikipedia's](#) [collaborative](#) [social](#) [network](#) [structure](#). In a paper presented earlier this year at [SIGCHI '11](#) authors discuss the results of a small series of semi-structured user interviews with Wikipedia administrators and editors. The results point to a number of elements in the design of Wikipedia talk pages, suggesting that editors find it hard to keep up-to-date with temporarily posted discussions that are other scattered across multiple pages. The interviews suggest that talk pages often become the target of support requests by the editors that are involved. The authors propose several design and specific actions to improve the way in which the main weaknesses of Wikipedia talk pages. In the remainder of the paper the authors introduce a lightweight solution to allow the affective categorization of comments posted on article talk pages to automatically structure them with an "ICE" mark-up. This mark-up can then be applied to end users with the aid of a classification tool, implemented and evaluated in [SIGCHI](#), and potentially used to generate granular notifications. In a poster presented last month at [MobiSys '11](#), the same team of researchers give an overview of such a project on [SIGCHI](#) discussions and features with a diagram the complexity of decision discussions and procedures in the English Wikipedia. [View on ArXiv](#)


The inaugural edition of the Wikimedia Research Newsletter, published on July 25, 2011.

## Open access vs. closed access publications

Complete references of the publications featured in the newsletter can be found at the bottom of each issue. Publications that are either self-archived in an open access repository or published in an open access journal will be marked with an *open access* icon next to the download link, e.g.:

Laniado, David, Riccardo Tasso, Y. Volkovich, and Andreas Kaltenbrunner. *When the Wikipedians talk: network and tree structure of Wikipedia discussion pages*. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 177-184, 2011. **PDF** <sup>[3]</sup> .

Publications that are not open access (i.e. behind a paywall or tied to institutional subscriptions) will be marked with a *closed access* icon:

Dalip, Daniel Hasan, Raquel Lara Santos, Diogo Rennó Oliveira, Valéria Freitas Amaral, Marcos André Gonçalves, Raquel Oliveira Prates, Raquel C.M. Minardi, and Jussara Marques de Almeida (2011). GreenWiki: A tool to support users' assessment of the quality of Wikipedia articles. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11)*, 469. New York, NY, USA: ACM Press. **DOI** <sup>[4]</sup> .

## Archives

- WRN 2(2) – February 2012
- WRN 2(1) – January 2012
- WRN 1(6) – December 2011
- WRN 1(5) – November 2011
- WRN 1(4) – October 2011
- WRN 1(3) – September 2011
- WRN 1(2) – August 2011
- WRN 1(1) – July 2011 (inaugural edition)
- Recent research – Signpost, 6 June 2011
- Recent research – Signpost, 11 April 2011

## Contact

For general queries on the research newsletter other than project or paper contributions you can leave a message on the talk page or mail us at: [researchnews@wikimedia.org](mailto:researchnews@wikimedia.org) <sup>[5]</sup>

## References

- [1] <http://lists.wikimedia.org/mailman/listinfo/wiki-research-l>
- [2] <http://etherpad.wikimedia.org/WRN201203>
- [3] <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2764/3301>
- [4] <http://dx.doi.org/10.1145/1998076.1998190>
- [5] <mailto:researchnews@wikimedia.org>

---

# Issue 1(1): July 2011

---

## Talk page interactions; Wikipedia at the Open Knowledge Conference; Summer of Research; brief notes

**With contributions by:** Junkie.dolphin, Lilaroja, HaeB, DarTar, Steven Walling, Daniel Mietchen.

This is the third overview of recent research published on Wikipedia and other Wikimedia projects featured in the Signpost (previous issues: June 6, April 11), and the inaugural issue of the Wikimedia Research Newsletter, intended to become a monthly feature. In addition to a focus on coverage of the research done by academics outside Wikimedia, this issue includes contributions funded by the Wikimedia Foundation. If you want your research to be featured in this monthly newsletter, you can tell us about your work by submitting it for consideration.

### **Edit wars and conflict metrics**

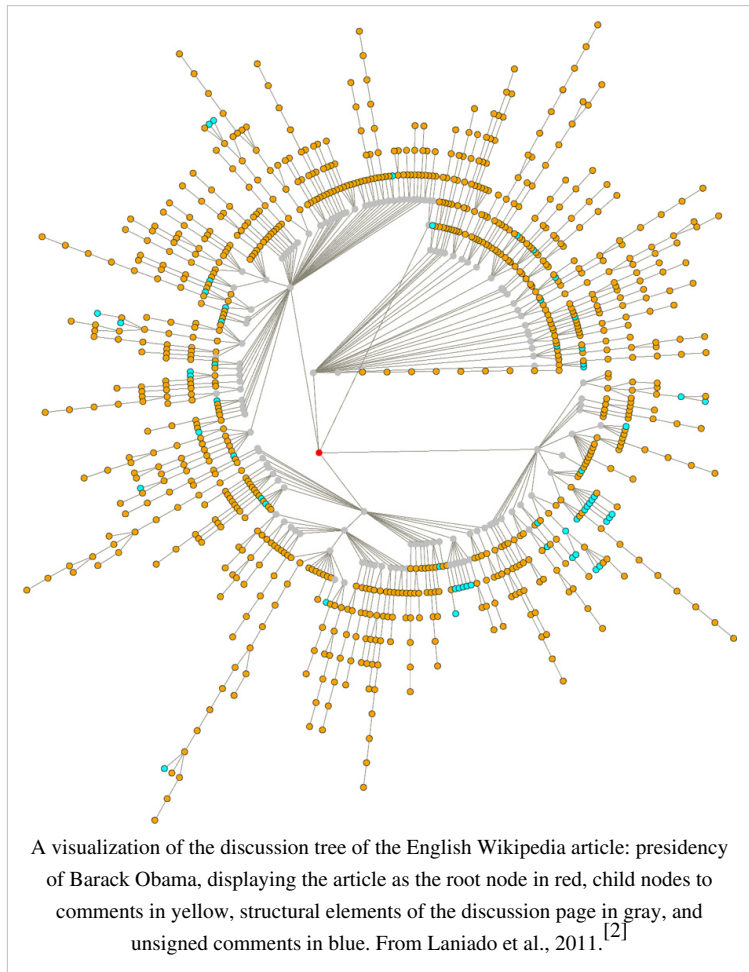
A study covered in the previous edition of the research newsletter was extended and published by the authors on *ArXiv*. The authors report a new method for classifying how disputed a Wikipedia article is, in the attempt to detect controversies and edit wars. At its core the method is based on looking at pairs of editors who have mutually reverted each other, and using their respective edit counts to define an overall metric of conflict. Even though this formula is not immediately intuitive, the authors motivate it by means of special diagrams called "revert maps" depicting such pairs of editors on the Cartesian space. In the second part of the study, the authors use this classifier to select two samples of pages, respectively of disputed and non-disputed topics, and analyze the time series of revisions to these pages; while they find that both time series are characterized by bursts of user activity, they claim that there is a qualitative difference between the two, although their analysis appears to lack any form of statistical hypothesis testing. They apply a priority-based model of editor activity that has been already proposed to explain human activity on the web, and find two distinctive patterns of activity that can help class "good" guys vs "bad" guys. <sup>[1]</sup>

## The anatomy of a Wikipedia talk page

Several works appeared over the last month focus on the structure and nature of social interaction supported by Wikipedia's discussion pages, both from a quantitative and qualitative perspective.

- **Wikipedia discussions shallow in geography and history, but deep in philosophy, law, language and beliefs.**

A study conducted by a team of researchers based in Milan and Barcelona and presented last week at ICWSM '11 <sup>[3]</sup> looks into the properties of the social interaction of participants in discussions held on talk pages.<sup>[2]</sup> The paper highlights a number of methodological issues in studying social network properties in Wikipedia. Social ties in Wikipedia are implicit, insofar as there is no representation of an explicit link between two Wikipedia users. A conversation between users allows to infer an implicit social network. Inferring such networks in Wikipedia, however, is challenging for two main reasons: because of the lack of structure of talk pages (which makes conversations hard to parse) and because of the fact that discussion threads can be very dispersed and take place in various locations (e.g. an article talk page plus a variable number of personal user talk pages). Despite these difficulties, the study succeeds at analyzing the properties of two types of *social networks centered on article discussions* (those happening on article talk pages and those that focus on an article but take place via replies on user talk pages) and a *user-centric social network* (i.e. the network defined by direct messages left by users on their talk pages). The three networks exhibit some interesting dissimilarities in terms of in- and out-degree of their nodes and in the proportion of overlap between their edges, suggesting that user-centered and article-centered communication are supported by substantially different networks. The study moves on to study the degree assortativity of these networks, or the tendency of users to create links with other users having a similar number of links. A striking difference emerges in the comparison with conversations in Slashdot, that are characterized by a strong assortativity, and discussion networks in Wikipedia, that display a systematic disassortativity, an indication of the specificity of social interactions that occur in Wikipedia compared to other social media. As the authors summarize, "Wikipedians who reply to many other users in article talk pages tend to interact mostly with users having few connections, i.e. newbies and inexperienced users, while the Wikipedians who receive replies from many users tend to interact preferentially with each other." The study moves on to consider the depth and popularity of article-centered discussions, and identifies metrics of the contentiousness of these discussions based on their depth and the number of mutual replies among users participating in the same thread. Finally, the study characterizes the size, frequency and structure of discussions across different article categories and finds that, although "Geography" and "History" account together for almost half of all discussions in the English Wikipedia, they tend to host shallow threads, whereas "Philosophy", "Law", "Language"



and those that focus on an article but take place via replies on user talk pages) and a *user-centric social network* (i.e. the network defined by direct messages left by users on their talk pages). The three networks exhibit some interesting dissimilarities in terms of in- and out-degree of their nodes and in the proportion of overlap between their edges, suggesting that user-centered and article-centered communication are supported by substantially different networks. The study moves on to study the degree assortativity of these networks, or the tendency of users to create links with other users having a similar number of links. A striking difference emerges in the comparison with conversations in Slashdot, that are characterized by a strong assortativity, and discussion networks in Wikipedia, that display a systematic disassortativity, an indication of the specificity of social interactions that occur in Wikipedia compared to other social media. As the authors summarize, "Wikipedians who reply to many other users in article talk pages tend to interact mostly with users having few connections, i.e. newbies and inexperienced users, while the Wikipedians who receive replies from many users tend to interact preferentially with each other." The study moves on to consider the depth and popularity of article-centered discussions, and identifies metrics of the contentiousness of these discussions based on their depth and the number of mutual replies among users participating in the same thread. Finally, the study characterizes the size, frequency and structure of discussions across different article categories and finds that, although "Geography" and "History" account together for almost half of all discussions in the English Wikipedia, they tend to host shallow threads, whereas "Philosophy", "Law", "Language"

and “Belief” are characterized by the deepest discussions and involve the largest amount of participants.

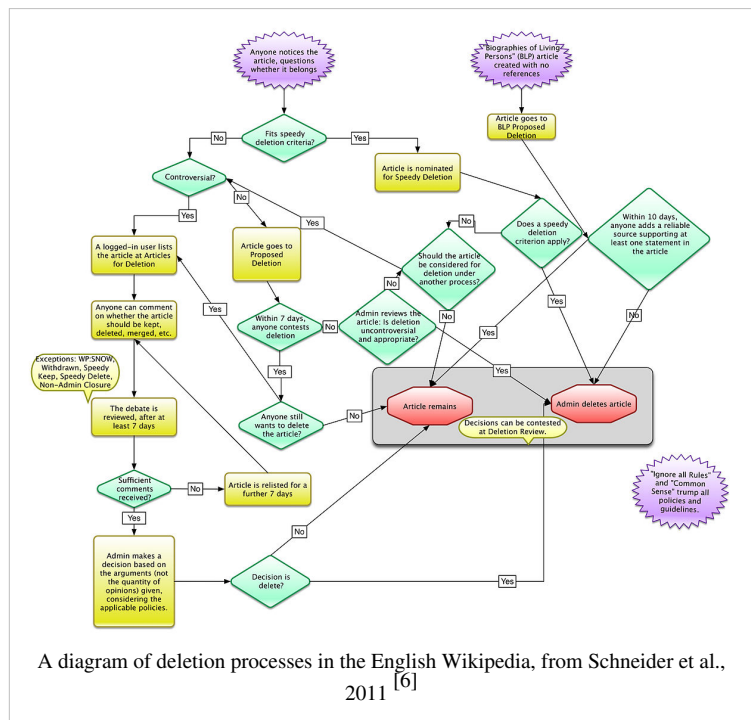
Two of the authors also gave a presentation at last month's Hypertext 2011 conference in Eindhoven, titled "Co-authorship 2.0: patterns of collaboration in Wikipedia [4]".

- **Building consensus in talk pages: authority and alignment.**

A group of researchers based at the University of Washington released an annotated corpus of discussions from Wikipedia talk pages encoding two types of social acts: alignment moves and authority claims.<sup>[5]</sup> In the authors' own words, "an *authority claim* is a statement made by a discussion participant aimed at bolstering their credibility in the discussion. An *alignment move* is a statement by a participant which explicitly positions them as agreeing or disagreeing with another participant or participants regarding a particular topic". Studying discussions with the lens of authority and alignment can help shed light on consensus-building strategies used by participants in Wikipedia discussions. The dataset also offers qualitative materials that, the authors submit, can be built upon to produce computational models of online debates. The data spans 365 discussions that occurred in 47 talk pages between 2002 and 2008, involving a total of 1,509 editors. After presenting the corpus, the study presents an analysis comparing editor activity metrics with the propensity of adopting one of the above social strategies. The authors introduce an editor's *v-index* (or *veteran index*) defined as the greatest *v* such that the editor has made at least *v* edits within the past *v* months and report that this indicator of editor activity positively correlates with the proportion of authority claims made in a discussion. The authors also observe that making an authority claim makes a user "significantly more likely to be the target of an alignment move within the subsequent 10 turns compared to turns that did not contain any claims".

- **Shortcomings in the design of Wikipedia talk pages.**

Researchers from the National University of Ireland, Galway presented some work in progress from a project aimed at understanding Wikipedia coordination spaces and costs. In a paper presented earlier this year at SAC '11 the authors discuss the results of a small series of semi-structured user interviews with Wikipedia administrators and editors.<sup>[7]</sup> The results point at a number of drawbacks in the design of Wikipedia talk pages, suggesting that editors find it hard to keep up-to-date with temporally sparse discussions that are often scattered across multiple pages. The interviews also suggest that talk pages often become the target of support requests by new editors that go unnoticed. The lack of connection between discussions and the article itself (e.g. links between threads and specific sections or topics of the article) also emerges as one of the main weaknesses of Wikipedia talk pages. In the remainder of the paper the authors introduce a lightweight solution to allow effective categorization of comments posted on article talk pages by semantically enriching them with an RDF markup. This markup can then be exposed to end users with the aid of a JavaScript bookmarklet, manipulated and exported via SPARQL, and potentially used to generate granular notifications. In a poster presented last month at *WebSci '11*, the same team of researchers gives an overview of work in progress on AfD discussions and illustrates with a diagram the complexity of deletion discussions and procedures in the English Wikipedia.<sup>[6]</sup>



In the remainder of the paper the authors introduce a lightweight solution to allow effective categorization of comments posted on article talk pages by semantically enriching them with an RDF markup. This markup can then be exposed to end users with the aid of a JavaScript bookmarklet, manipulated and exported via SPARQL, and potentially used to generate granular notifications. In a poster presented last month at *WebSci '11*, the same team of researchers gives an overview of work in progress on AfD discussions and illustrates with a diagram the complexity of deletion discussions and procedures in the English Wikipedia.<sup>[6]</sup>

## Wikipedians as "Janitors of Knowledge"

In a paper titled "Janitors of Knowledge: Constructing Knowledge in the Everyday Life of Wikipedia Editors",<sup>[8]</sup> researcher Olof Sundin of Lund University applies concepts from Science and technology studies to an online ethnography study of the Swedish Wikipedia community, focusing on the role of references in particular.

He conducted interviews with eleven active users of the Swedish Wikipedia (out of 20 contacted via e-mail) who had given "informed consent according to the recommendations of the Swedish Research Council". Their activity, as well as discussion on the village pump and on the talk pages of some articles, were observed from August 2009 to February 2010. (The paper does not link diffs of the users' comments, due to privacy reasons.) They were between 20 and 50 years old, with diverse jobs and outside interests. Among other observations, the paper states that "For most of the informants the watch-list ... is the starting point for their [everyday] activities", and that Wikipedia is also a place for identity construction, .... For Wikipedia editors, to edit is not just something you do, it is also a part of who you are". The title refers to the finding that "Cleaning work [e.g. reverting vandals] seems to be the central activity for almost all of the participants" of the interviews. The informants state that citing references has become more important on Wikipedia in recent years, also evidenced by the introduction (in November 2009) of a requirement to cite at least one reference in the criteria for inclusion of new articles in a "New Written Articles of the Week" page (similar to the English Wikipedia's Did You Know). One section is devoted to Wikipedia's "hierarchy of references" (by reliability), mentioning the Swedish Wikipedia's equivalent of w:WP:RS.

As theoretical framework, Sundin uses an actor-network theory interpretation of Wikipedia, which he explains as follows: "Within such a perspective, the editors, form and functions, core policies, guidelines of Wikipedia, its millions of articles and discussions, references, and users around the world can all be seen as actors, as they make each other do something; they construct, uphold and transform Wikipedia as we know it. An actor, for instance a functional feature in Wikipedia called the watch-list, that makes it easier for the editors to scan new contributions, or a policy document, makes other actors act in a particular way. ... Some actors have a more central role than others and some of these, if we draw on Callon (1986), are so central that they can be called obligatory passage points. An obligatory passage point can be thought of as a threshold that other actors need to pass or adjust to." As such an obligatory passage point in Wikipedia's network of actors, Sundin identifies the Verifiability policy.

## Use of Wikipedia among law students: a survey

An article in *The Law Teacher* titled: "Embracing Wikipedia as a research tool for law: to Wikipedia or not to Wikipedia?" describes an anonymous survey among 101 Australian students (30 senior secondary high school students enrolled in legal studies, and 71 law degree students in their first and second year at the University of Southern Queensland) about their use and perceptions of Wikipedia.<sup>[9]</sup> Their results indicate "that the majority (78%) of all students surveyed are currently using Wikipedia for some form of legal (30%) or other research (37%) or as a source of general information (11%)." One of the 101 students admitted to have vandalized Wikipedia articles, while two stated to have corrected errors in Wikipedia. The use of Wikipedia for legal research among the first-year university students was much lower than among the high-school students, which the authors conjecture is "a result of legal research skills training and warnings against its use, and perhaps even a result of cultural adaptation. Seventy-eight percent of the first year law students surveyed acknowledged that Wikipedia can be unreliable and/or inaccurate." However, Wikipedia usage for legal or other research increased again for the second year university students, which the authors surmise could have to do with the students becoming "a little more streetwise within the university context and [finding] the convenience of Wikipedia appealing." Apart from the poll results, the paper contains a small literature survey about "Wikipedia as a teaching and learning resource", observing that "the use of wikis in legal education is in its infancy. Several of the case studies in the literature reported positive outcomes," and qualitative results from an "informal preliminary investigation into academic perceptions of Wikipedia as a research source in law" ("All the academics consulted considered Wikipedia an unreliable source for legal information ... Some acknowledged a role for Wikipedia as a source for legal or incidental background information" with



qualifications about accuracy and reliability). Still, "the authors argue that using Wikipedia as a tertiary source for assimilating broad overview information, for both legal and incidental research, to define and identify keywords for further research, and as a link to other resources, is acceptable when the issues surrounding the discerning use of any secondary source, peer reviewed or not, are fully understood", and that "Academics can and should contribute to Wikipedia either directly, through the contribution of research, or indirectly, through the mentoring of student contributions which can be incorporated into course content and assessments." Among other conclusions, the authors suggest "encouraging universities to develop policies consistent with academic contribution to Wikipedia".

## Miscellaneous

- **Turning back Wikipedia's clock:** In a paper titled "Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History",<sup>[10]</sup> three researchers from the Darmstadt University of Technology present a software that allows easy access of the state of Wikipedia corresponding to a particular point in time, both for single article revisions and for whole history dumps up to that moment. As motivation, they note that large-scale access of single revisions via Wikipedia's own API is inefficient since data needs to be transmitted over the Internet, whereas the downloadable XML dumps provided by the Wikimedia Foundation are in a format that doesn't allow easy access of single revisions. They emphasize the importance of these dumps for Natural Language Processing analyses of Wikipedia, and that the reproducibility of such research is jeopardized by the fact that "older snapshots [are] becoming unavailable as there is no official backup server." The authors' solution is realized as an extension of the existing Java Wikipedia Library (JWPL)<sup>[11]</sup>. To store the dump in a format that allows fast access to revisions but still saves space, they developed their own diff algorithm based on a longest common substring search.
- On his personal blog, Paolo Massa (w>User:Phauly) gave a "Report of ACM Hypertext 2011 conference"<sup>[12]</sup> from the perspective of a Wikipedia researcher.
- The inaugural issue<sup>[13]</sup> of *Critical Studies in Peer Production*<sup>[14]</sup>, a new open access academic journal, published an article<sup>[15]</sup> by Mathieu O'Neil titled: "The sociology of critique in Wikipedia". All contents from this journal are CC-BY licensed.
- A blog posting titled "Who writes Wikipedia? An information-theoretic analysis of anonymity and vandalism in user-generated content"<sup>[16]</sup> referenced a widely cited study by Aaron Swartz (see also this week's Wikipedia Signpost "In the news"), who in 2006 found that anonymous users contributed much more of Wikipedia's content than the core of registered users. Instead of examining the text that survived to the current version, the blogger looked at reverted/unreverted edits as a crude measure of quality, and instead of counting edits, measured "the information-theoretic gain in each revision", as measured by LZMA compression. For performance reasons, the analysis was restricted to pages starting with the letter "M". Among various other findings, the post states that "Registered users dipped to contributing as much vandalism as content in 2007, and have taken an upswing to over three times as much good content. Anonymous users dipped to contributing as much vandalism as content in 2005, and through 2010 are contributing roughly twice as much vandalism as content".
- **Using expertise credits from Citizendium to recommend Wikipedia articles:** An article in this month's issue of the "Journal of Information Processing", titled "Classification of Recommender Expertise in the Wikipedia Recommender System"<sup>[17]</sup>, reports improvements in the existing "Wikipedia Recommender System", a "collaborative filtering system with trust metrics, i.e., it provides a rating of articles which emphasizes feedback from recommenders that the user has agreed with in the past", by considering the recommenders' areas of expertise. To determine these areas, the paper uses the self-reported expertise areas that Citizendium contributors have to state when signing up for one of that online encyclopedia's topic work groups.
- A team of Brazilian researchers from Universidade Federal de Minas Gerais presented at JCDL '11<sup>[18]</sup> a tool called *GreenWiki*, designed to "help improve users awareness about the quality of a Wikipedia article as well as their assessment of it".<sup>[19]</sup>

- An article in last month's issue of *Information Research* examined "The search queries that took Australian Internet users to Wikipedia<sup>[20]</sup>". The results "suggest that Wikipedia is used more for lighter topics than for those of a more academic or serious nature. Significant differences among the various lifestyle segments were observed in the use of Wikipedia for queries on popular culture, cultural practice and science".

## Wikipedia research at OKCon 2011

On June 30 - July 1, the Open Knowledge Foundation held their annual meeting, the Open Knowledge conference (OKCon), this time in Berlin. On the first day, a workshop on Wikipedia & Research<sup>[21]</sup> took place, organized by Mayo Fuster Morell (member of the Research Committee of the Wikimedia Foundation), who agreed to report back for the *Signpost*.

A message was already sent by the simple observation that the room was packed with around 50 people, some of them even sitting on the floor. In a tweet<sup>[22]</sup>, Philipp Schmidt from P2P University commented: "*wikipedia research community growing and diversifying. I remember meetings with 5 people, now the room is packed. Great!*". The attendance at the workshop is a sign that there is high interest in the question of promoting research around Wikipedia. Furthermore, the good response could be seen from a double perspective: because addressing the questions is considered as important *per se*, but also in terms of good timing - a question of the right moment.

Since 2005, there has been an increasing interest within the scientific community in researching Wikipedia. In 2011, ten years after Wikipedia started, research on Wikipedia keeps growing, with a body of research and a community of researchers in place. In this regard, according to a recent review, there is currently a total of 2,100 peer-reviewed articles and 38 doctoral theses related to Wikipedia. The willingness to collaborate, to make use of synergies between research initiatives of various kinds, and to continue innovating (in what is already constituting one of the leading nodes of methodological innovation) have also increased and continue to mature. It seems that in 2011 and the coming years, we will see not only the continuation in terms of a quantitative increase, but also a qualitative jump towards a more organized and challenging stage of research initiatives from and around Wikipedia. This can be expected to translate into important changes at the research level, and the initiative of research being promoted by Wikipedia (not only about Wikipedia) is likely to be well received.

During the workshop, Mathias Schindler (from Wikimedia Deutschland) presented the RENDER project - a research project looking at knowledge diversity, which is the first experience of a Wikimedia Chapter engaging in a large research projects with other research partners at the European level.

Mayo Fuster Morell presented how Wikipedia had evolved over the years. Starting with quantitative analyses of large data sets and on the English version of Wikipedia as the predominant approach in early empirical research on Wikipedia, the focus then expanded to conducting research on other language versions, covering a larger variety of issues, such as socio-political questions, and also adopting qualitative methods. She also presented the Research Committee, a committee created by the WMF staff consisting of Wikimedia volunteers, researchers, and Wikimedia Foundation staff with the mandate to help organize policies, practices and priorities around Wikimedia-related research).

Daniel Mietchen (likewise a member of the Research Committee of the WMF) presented the draft for an open access and open data policy of the WMF as a requirement for research projects receiving significant WMF support.

Benjamin Mako Hill (Advisory Board member and intellectual property researcher at MIT, among others) was also present, but stepped back from his planned intervention in favor of allowing time for debate. During the discussion, the question of open data was the central theme of interest to the floor. Other than that, interest was also expressed in the question of data repositories.

The schedule was tight, and the session ended well before the discussions could have reached a conclusion. It remains clear that a continuation of the discussion is needed as much as occasions to meet and develop things together around Wikipedia research and promoting another way of doing research.

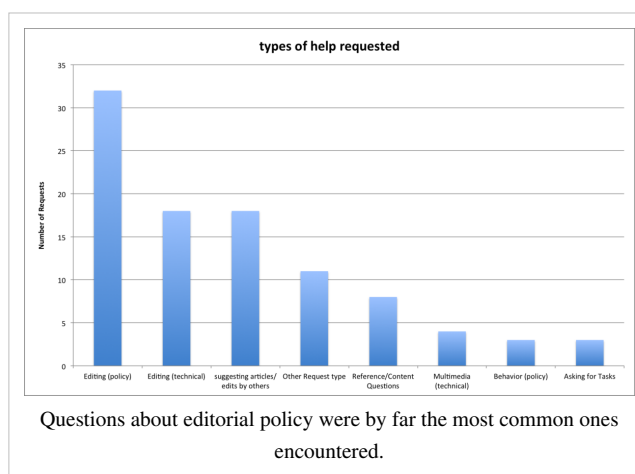
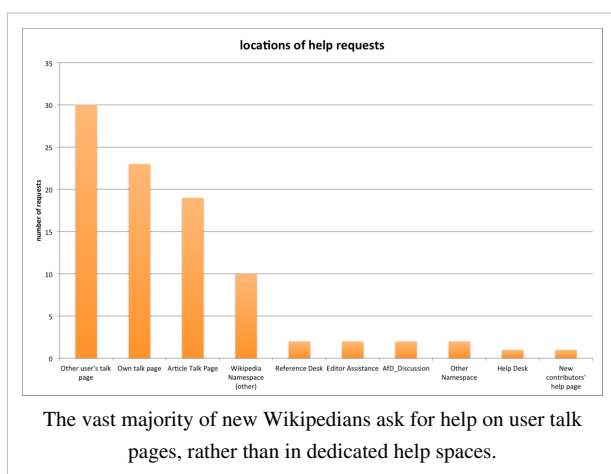
## Wikimedia Summer of Research: Three of the topics covered so far

The "Wikimedia Summer of Research" (WSoR, see previous coverage in the *Wikipedia Signpost*) is a three-month program (ending in September) sponsored by the Wikimedia Foundation which has brought together a team of eight academics who are working in the Foundation's Community Department. The goal of this program is to study the dynamics of the editing community, starting with English and focusing particularly on what factors can measurably be said to affect the decline in new editors. The following is a short look at three of the many areas studied so far. Other research can be found on Meta and on Commons.

### How New English Wikipedians Ask for Help

The early weeks of research by Jonathan Morgan, R. Stuart Geiger, and Shawn Walker were focused on how new editors find and interact with help spaces, both in the Help namespace and outside it. A combination of qualitative and quantitative methods have been used to address this issue, but the primary data was gathered through qualitative coding of randomized samples of new editors.

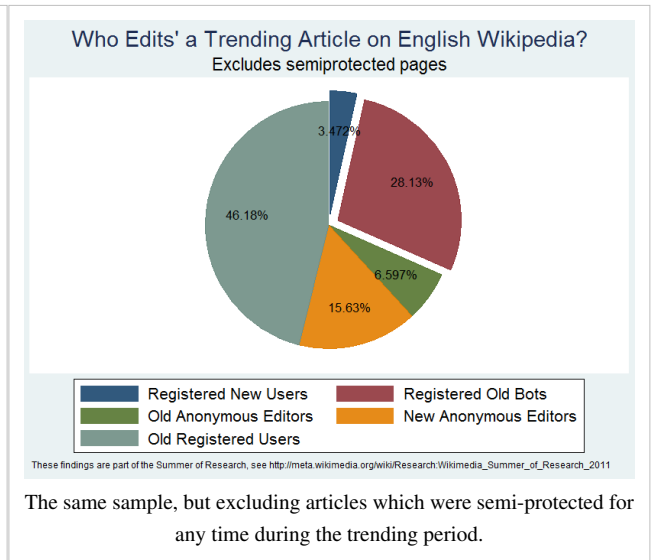
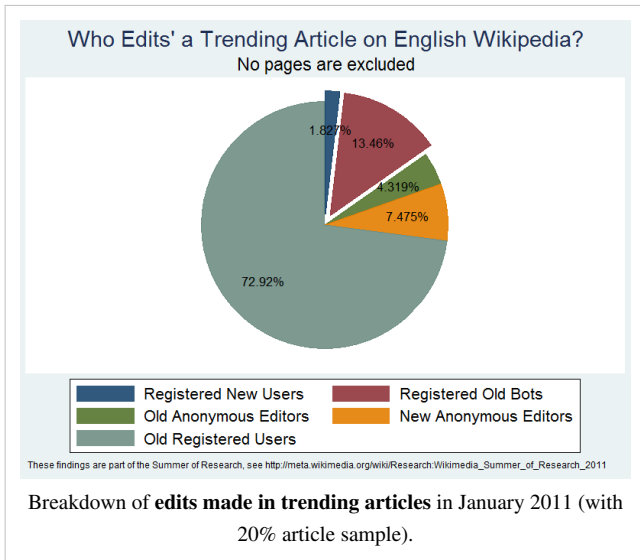
The following two charts were derived from the coding of activities by 445 new Wikipedians distributed from 2009-2011.<sup>[23]</sup>



### Who Edits Trending Articles on the English Wikipedia

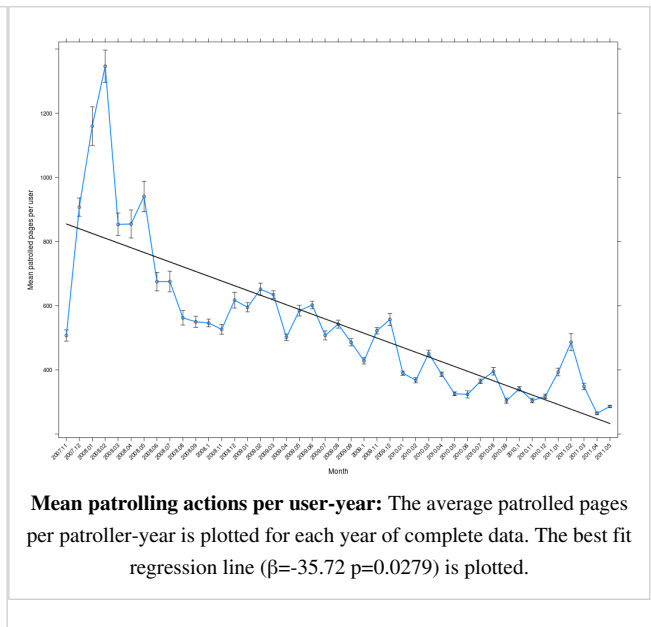
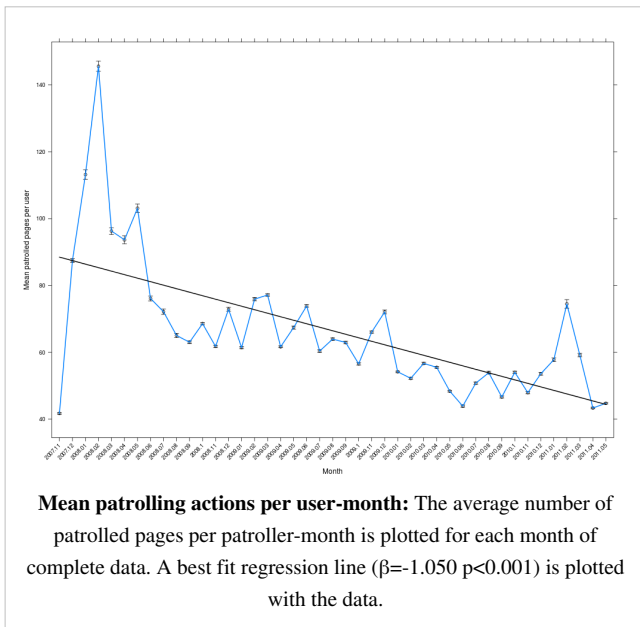
One question that was directed at the summer research team was whether trending articles — such as those about breaking current events or included in "In the news" on the Main Page — attracted a significant number of new editors compared to articles not affected by current events. Adjacent questions were whether those new editors who registered because interest in current events articles were more or less likely to become repeat editors of the encyclopedia.

Using a quantitative sample of a random 20% of the thousands of articles which were trending (in terms of traffic stats) in January 2011, this study by Yusuke Matsubara showed that, perhaps surprisingly, the number of newly registered editors who participate in current events articles is proportionally quite low.<sup>[24]</sup> However, the amount of participation from anonymous editors was more significant regardless of semi-protection. This hints that there may be an opportunity to invite good faith anonymous contributors on trending articles to participate further by registering accounts.



### The Workload of New Page Patrollers & Vandalfighters

One of the theories that has been proposed about the decline in participation by new editors is that newbie biting has increased over the years because more of the burden of policing vandalism, spam, etc. has been shouldered by fewer and fewer active New Page Patrollers and vandalfighters, which also contributes to burn out. To test this theory, summer researcher Aaron Halfaker looked at the work load of new page patrollers<sup>[25]</sup> and vandalfighters<sup>[26]</sup> since 2007 overall. It found that, like many things in Wikipedia, the trends follow a power law where the top contributors do most of the work. However, contrary to the hypothesis, the number of patrolling actions per editor (by both month and year) has been decreasing steadily.



## References

- [1] Sumi, R., T. Yasseri, A. Rung, A. Kornai, and J. Kertész (2011). Edit wars in wikipedia. ArXiv (<http://arxiv.org/abs/1107.3689>) [stat.ML]. **PDF** (<http://arxiv.org/pdf/1107.3689v1>) Open access
- [2] Laniado, David, Riccardo Tasso, Y. Volkovich, and Andreas Kaltenbrunner. *When the Wikipedians talk: network and tree structure of Wikipedia discussion pages*. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 177-184, 2011. **PDF** (<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2764/3301>) Open access.
- [3] <http://www.aaai.org/Library/ICWSM/icwsml1contents.php>
- [4] <http://airlab.elet.polimi.it/images/2/23/Coauthorship2.pdf>
- [5] Bender, E.M., J.T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf (2011). Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 48-57. **PDF** (<http://acl.eldoc.ub.rug.nl/mirror/W/W11/W11-0707.pdf>) Open access
- [6] Schneider, Jodi, and Alexandre Passant (2011). Arguments about Deletion: Guiding New Users in Making Good Arguments. In *Proceedings of the 2011 ACM Web Science Conference (WebSci '11)*. **PDF** ([http://www.websci11.org/fileadmin/websci/Posters/187\\_paper.pdf](http://www.websci11.org/fileadmin/websci/Posters/187_paper.pdf)) Open access
- [7] Schneider, Jodi, Alexandre Passant, and John G. Breslin (2011). Understanding and improving Wikipedia article discussion spaces. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC '11)*, 808. New York, NY: ACM Press. **DOI** (<http://dx.doi.org/10.1145/1982185.1982358>) • **PDF** (<http://jodischneider.com/pubs/sac2011.pdf>) Open access
- [8] Sundin, Olof (2011) Janitors of Knowledge: Constructing Knowledge in the Everyday Life of Wikipedia Editors. *Journal of Documentation*, 67, no. 5: 6. (abstract) (<http://www.emeraldinsight.com/journals.htm?issn=0022-0418&volume=67&issue=5&articleid=1931285&show=html>). Closed access
- [9] Barnett, Eola, and Roslyn Baer (2011), Embracing Wikipedia as a research tool for law: to Wikipedia or not to Wikipedia? *The Law Teacher* 45, no. 2: 194-213. **DOI** (<http://dx.doi.org/10.1080/03069400.2011.578883>). Closed access
- [10] Ferschke, Oliver, Torsten Zesch, and Iryna Gurevych (2011) Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11)*, 97-102, Association for Computational Linguistics. **PDF** (<http://www.aclweb.org/anthology/P/P11/P11-4017.pdf>) Open access
- [11] <http://code.google.com/p/jwpl/>
- [12] [http://www.gnuband.org/2011/06/17/report\\_of\\_acm\\_hypertext\\_2011\\_conference/](http://www.gnuband.org/2011/06/17/report_of_acm_hypertext_2011_conference/)
- [13] [http://listcultures.org/pipermail/cpov\\_listcultures.org/2011-June/000354.html](http://listcultures.org/pipermail/cpov_listcultures.org/2011-June/000354.html)
- [14] <http://cspp.oekonux.org>
- [15] <http://cspp.oekonux.org/research/mass-peer-activism/rs-1.2-sociology-of-critique>
- [16] <http://slightlynew.blogspot.com/2011/05/who-writes-wikipedia-information.html>
- [17] <http://joi.jlc.jst.go.jp/JST.JSTAGE/ipsjjip/19.345?from=CrossRef>
- [18] <http://jcdl2011.org/>
- [19] Dalip, Daniel Hasan, Raquel Lara Santos, Diogo Rennó Oliveira, Valéria Freitas Amaral, Marcos André Gonçalves, Raquel Oliveira Prates, Raquel C.M. Minardi, and Jussara Marques de Almeida (2011). GreenWiki: A tool to support users' assessment of the quality of Wikipedia articles. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11)*, 469. New York, NY, USA: ACM Press. **DOI** (<http://dx.doi.org/10.1145/1998076.1998190>) Closed access
- [20] <http://informationr.net/ir/16-2/paper476.html>
- [21] <http://okcon.org/2011/programme/wikipedia-research-the-innovative-character-of-wikipedia-research-and-the-new-challenges-and-opportunities-associated-with-it>
- [22] <http://twitter.com/sharingnicely/status/86409043664584704>
- [23] Morgan, J. T., Geiger, R.S., Pinchuk, M. and Walker, S. (2011), New Users and Help: When, Where, Why and How. Open access
- [24] Matsubara, Y. (2011), Research:Trending articles and new editors. Open access
- [25] Halfaker, A. (2011), Research:Patroller work load. Open access
- [26] Halfaker, A. (2011), Research:Vandal fighter work load. Open access

# Issue 1(2): August 2011

## Article promotion by collaboration; deleted revisions; Wikipedia's use of open access; readers unimpressed by FAs; swine flu anxiety

With contributions by: DarTar, Mietchen and Tbayer

### Effective collaboration leads to earlier article promotion

A team of researchers from the MIT Center for Collective Intelligence investigated the structure of collaboration networks in Wikipedia and their performance in bringing articles to higher levels of quality. The study<sup>[1]</sup>, presented at *HyperText 2011*<sup>[2]</sup>, defines collaboration networks by considering editors who edited the same article and exchanged at least one message on their respective talk pages. The authors studied whether a pre-existing collaboration network or structured collaboration management via WikiProjects accelerate the process of quality promotion of English Wikipedia articles. The metric used is the time it takes to bring articles from B-class to GA or from GA to FA on the Wikipedia 1.0 quality assessment scale. The results show that the WikiProject importance of an article increases its promotion rate to GA and FA by 27% and 20%, respectively. On the other hand, the number of WikiProjects an article is part of reduces the rate of promotion to FA by 32%, an effect that the authors speculated could imply that these articles are broader in scope than those claimed by fewer WikiProjects. Pre-existing collaboration also dramatically affects the rate of promotion to GA and FA (with 150% and 130% increases, respectively): prior collaborative ties significantly accelerate article quality promotion. The authors also identify contrasting effects of network structure (cohesiveness and degree centrality) on the increase of GA and FA promotion times.

### Deleted revisions in the English Wikipedia

Andrew G. West and Insup Lee from the University of Pennsylvania conducted the first-ever study examining revision deletion in the English Wikipedia,<sup>[3]</sup> in a paper to be presented at the upcoming *WikiSym 2011*<sup>[4]</sup> symposium. Several scholarly works have studied standard deletion procedures in Wikipedia; this paper presents original results on "contributions that are not simply *undone* but deleted from revision histories and public views". Revision deletion, or redaction, is a process enabled by a feature (RevDelete)

introduced in 2009 for the purpose of removing *dangerous contents*, such as user contributions infringing copyright or inserting defamation, insults, or individual privacy threats. Access to this feature was initially restricted to users with *oversight* privileges and later extended to *administrators*. The study analyzes a year of public deletion logs and the contents of deleted revisions, by comparing two snapshots of edits data from the English Wikipedia. The authors identify 49,161 unique deleted revisions produced by 18,907 incidents. The number of deleted revisions is higher than the number of incidents, as some categories of dangerous content survive for more than a single revision and

**Selected revision of [Wikipedia:Test page](#):**

- [\(diff\) 01:00, June 01, 2009 Example User](#) *(Adding something to this page)*

**Deleted revisions and events will still appear in the page history and logs, but parts of their content will be inaccessible to the public.**

Please confirm that you intend to do this, that you understand the consequences, and that you are doing this in accordance with [the policy](#).

Set visibility restrictions

Delete revision text

Delete edit comment

Delete editor's username/IP

**Suppress data from administrators as well as others**

Log comment:

The RevDelete dialog.

their deletion consequently affects a series of revisions. By analyzing the reasons for deletion in the deletion log, the authors find offensive content directed at specific individuals to be the most frequent cause of deletion incidents (58%), followed by acts of disruption (29%), and copyright infringement (11%). Results for incidents that occurred after May 2010 indicate that the two-hour median detection interval calculated on all incidents increases to 21.6 days for copyright-related incidents, suggesting that the latter are much harder to detect. For the same reason, copyright-related incidents span longer series of deleted revisions (12.5 on average, whereas 89% of all incidents result in a single deleted revision). Considering the amount of time that subsequently deleted contents remained visible on a page, the authors find that the median of 2 minutes (calculated over all incidents) increases to 21 days in the case of copyright incidents (virtually the same time of their detection interval). The study reports that at least 0.05% of revisions made in 2010 contained dangerous contents and that 0.007% of all page views in 2010 resulted in the temporary exposure of these contents.

## Wikipedia and open-access repositories

The paper "Wikipedia and institutional repositories: an academic symbiosis?"<sup>[5]</sup> is concerned with Wikipedia articles citing primary sources when suitable secondary ones (as per WP:SCHOLARSHIP) are not available. Only about 10% of scholarly papers are published as open access, but another 10% are freely available through self-archiving, thus doubling in practice the number of scholarly primary resources that Wikipedia editors have at their disposal. The article describes a sample of institutional repositories from the major higher-education institutions in New Zealand, along with three Australian institutions serving as controls, and analyses the extent to which they are linked from Wikipedia (across languages).



The so far only illustration in the Stirling Gardens article is a retouched version of an image found in an institutional repository.

Using Yahoo! Site Explorer, a total of 297 links were estimated to go from Wikipedia articles to these repositories (40% of which went to the three Australian controls), mostly to support specific claims but also (in 35% of the cases) for background information. In terms of document type linked from Wikipedia, PhD theses, academic journal articles and conference papers each scored about 20% of the entries, whereas in terms of Wikipedia language, 35% of links came from non-English Wikipedias.

The paper cites strong criticism of institutional repositories<sup>[6]</sup> but proposes "a potential symbiosis between Wikipedia and academic research in institutional repositories" – Wikipedia getting access to primary sources, and institutional repositories growing their user base – as a new reason that "academics should be systematically placing their research work in institutional repositories". Ironically, the author himself did not<sup>[7]</sup> follow this advice. However, such potential alignments between Wikimedians and open access have been observed in related contexts – according to the expert participation survey. For instance, Wikipedia contributors are more likely<sup>[8]</sup> to have most or all of their publications freely available on the web.

As is custom in academia, the paper does not provide links to the underlying data, but the Yahoo! Site Explorer queries can be reconstructed<sup>[9]</sup> (archived example<sup>[10]</sup>) or compared to Wikipedia search results<sup>[11]</sup> and site-specific Google searches<sup>[12]</sup>. There is also code from linkypedia<sup>[13]</sup> and the Wikipedia part of the PLoS Altmetrics study<sup>[14]</sup> that could both be adapted for automating such searches.

## Quality of featured articles doesn't always impress readers

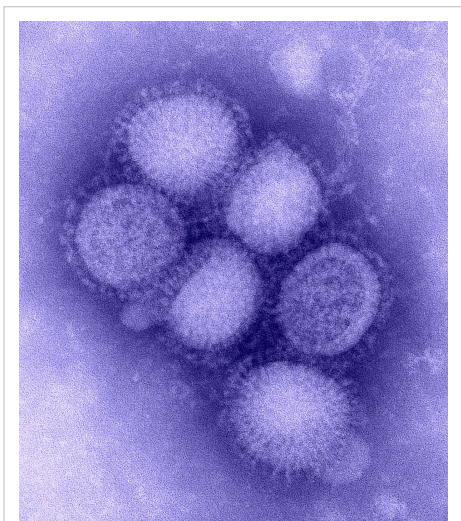
In an article titled "Information quality assessment of community generated content: A user study of Wikipedia" (abstract <sup>[15]</sup>),<sup>[16]</sup> published this month by the *Journal of Information Science*, three researchers from Bar-Ilan University reported on a study examining judgment of Wikipedia's quality by non-expert readers (done as part of the 2008 doctoral thesis of one of the authors, Eti Yaari, which was already covered <sup>[17]</sup> in Haaretz at the time).

The paper starts with a review of existing literature on information quality in general, and on measuring the quality of Wikipedia articles in particular. The authors then describe the setup of their qualitative study: 64 undergraduate and graduate students were each asked to examine five articles from the Hebrew Wikipedia, as well as their revision histories (an explanation was given), and judge their quality by choosing the articles they considered best and worst. The five articles were pre-selected to include one containing each of four quality/maintenance templates used on the Hebrew Wikipedia: featured, expand, cleanup and rewrite, plus one "regular" article. But only half of the participants were shown the articles with the templates. Participants were asked to "think aloud" and explain their choices; the audio recording of each session (on average 58 minutes long) was fully transcribed for further analysis, which found that the criteria mentioned by the students could be divided into "measurable" criteria "that can be objectively and reliably assigned by a computer program without human intervention (e.g. the number of words in an article or the existence of images)" and "non-measurable" ones ("e.g. structure, relevance of links, writing style", but also in some cases the nicknames of the Wikipedians in the version history). Interestingly, a high number of edits was both seen as a criterion indicating good quality by some, and indicating bad quality by others, and likewise for a low number of edits.

Comparing the quality judgments of the study's participants with that of Wikipedians as expressed in the templates revealed some striking differences: "The perceptions of our users regarding quality did not always coincide with the perceptions of Wikipedia editors, since in fewer than half of the cases the featured article was chosen as best". In fact, "in three cases, the featured article was chosen as the lowest quality article out of the five articles assessed by these participants." However, those participants who were able to see the templates chose the "featured" article considerably more often as the best one, "even though the participants did not acknowledge the influence of these templates".

## In swine flu outbreak, Wikipedia reading preceded blogging and newspaper writing

A paper published in *Health Communication* examined "Public Anxiety and Information Seeking Following the H1N1 Outbreak" (of swine influenza in 2009) by tracking, among other measures, page view numbers on Wikipedia, which it described as "a popular health reference website" (citing a 2009 paper co-authored by Wikipedian Tim Vickers: "Seeking health information online: Does Wikipedia matter?"<sup>[18]</sup>). Specifically, the researchers - psychologists from the University of Texas and the University of Auckland - selected 39 articles related to swine flu (for example, H1N1, hand sanitizer, and fatigue) and examined their daily page views from two weeks prior to two weeks after the first announcement of the H1N1 outbreak during the 2009 flu pandemic. (The exact source of the page view numbers is not stated, but a popular site <sup>[19]</sup> providing such data exists.) Controlling for variations per day of the week, they found that "the increase in visits to Wikipedia pages happened within days of news of the outbreak and returned to baseline within a few weeks. The rise in number of visits in response to the epidemic was greater the first week than the second week .... At its peak, the seventh day, there were 11.94 times as many visits per article on average."



Generated public anxiety in 2009, measurable in Wikipedia pageviews: The H1N1 influenza virus



While these findings may not be particularly surprising to Wikipedians who are used to current events driving attention for articles, the authors offer intriguing comparisons to the two other measures of public health anxiety they study in the paper: The number of newspaper articles mentioning the disease or the virus, and the number of blog entries mentioning the disease. "Increased attention to H1N1 happens most rapidly in Wikipedia page views, then in the blogs, and finally in newspapers. The duration of peak attention to H1N1 is shortest for the blog writers, followed by Wikipedia viewers, and is longest in newspapers." Examining correlations, they found that "The number of blog entries was most strongly related to the number of newspaper articles and number of Wikipedia visits on the same day. The number of Wikipedia visits was most strongly related to the number of newspaper articles the following day. In other words, public reaction is visible in online information seeking before it is visible in the amount of newspaper coverage." Finally, the authors emphasize the advantages of their approach to measure public anxiety in such situations over traditional approaches. Specifically, they point out that in the H1N1 case the first random telephone survey was conducted only two weeks after the outbreak, and therefore underestimated the initial public anxiety levels, as the author argue based on their combined data including Wikipedia pageviews.

### **Extensive analysis of gender gap in Wikipedia to be presented at WikiSym 2011**

A paper by researchers of GroupLens Research to be presented at the upcoming *WikiSym 2011* <sup>[4]</sup> symposium offers the most comprehensive analysis of gender imbalance in Wikipedia so far.<sup>[20]</sup> This study was covered by a summary in the August 15 *Signpost* edition and facilitated by a press release <sup>[21]</sup>, it generated considerable media attention. Below are some of the main highlights from this study:

- reliably tracking gender in Wikipedia is complicated due to the different (and potentially inconsistent) ways in which users can specify gender information.
- self-identified females comprise 16.1% of editors who started editing in 2009 and specified their gender, but they only account for 9% of the total number of edits by this cohort and the gap is even wider among highly active editors.
- the gender gap has remained fairly constant since 2005
- gender differences emerge when considering areas of contribution, with a greater concentration of women in the People and Arts areas.
- male and female editors edit user-centric namespaces differently: on average, a female makes a significantly higher concentration of her edits in the User and User Talk namespaces, mostly at the cost of fewer edits in Main and Talk.
- a significantly higher proportion of females have participated in the "Adopt a User" program as mentees.
- female editors have an overall lower probability of becoming admins. However, when controlling for experience measured by number of edits it turns out that women are significantly more likely to become administrators than their male counterparts.
- articles that have a higher concentration of female editorship are more likely to be contentious (when measured by proportion of edit-protected articles) than those with more males.
- in their very initial contributions, female editors are more likely to be reverted than male editors but there is hardly any statistical difference between females and males in how often they are reverted after their seventh edit. The likelihood of the departure of a female editor, however, is not affected more than that of a male by reverts of edits that are genuine contributions (i.e. not considered vandalism).
- females are significantly more likely to be reverted for vandalizing Wikipedia's articles and while males and females are temporarily blocked at similar rates, females are significantly more likely to be blocked permanently. In these cases, though, self-reported gender may be less reliable.

A second, unpublished paper addressing gender imbalance in Wikipedia ("Gender differences in Wikipedia editing") by Judd Antin and collaborators will be presented at *WikiSym 2011*.

## "Bandwagon effect" spurs wiki adoption among Chinese-speaking users

In a paper titled "The Behavior of Wiki Users"<sup>[22]</sup>, appearing in *Social Behavior and Personality: An International Journal*, two researchers from Taiwan used the Unified Theory of Acceptance and Use of Technology (UTAUT) "to explain why people use wikis", based on an online questionnaire distributed in July 2010 in various venues and to Wikipedians in particular. According to an online version of the article <sup>[23]</sup>, the survey generated 243 valid responses from the Chinese-speaking world, which showed that – similar to previous results for other technologies – each of the following "had a positive impact on the intention to use wikis":

- Performance expectancy (measured by agreement to statements such as "Wikis, for example Wikipedia, help me with knowledge sharing and searches")
- Effort expectancy (e.g. "Wikis are easier to use than other word processors.")
- Facilitating conditions (e.g. "Other wiki users can help me solve technical problems.")
- User involvement (e.g. "Collaboration on wikis is exciting to me.")

The impact of user involvement was the most significant. Social influence (e.g. "The people around me use wikis") was not found to play a significant role. On the other hand, the researchers state that a person's general susceptibility to the "bandwagon effect" (measured by statements such as "I often follow others' suggestions") "can intensify the impact of [an already present] intention to use wikis on the actual use ... This can be explained in that users tend to translate their intention to use into actual usage when their inclination receives positive cues, but the intention alone is not sufficient for them to turn intention into action. ... people tend to be more active in using new technology when social cues exist. This is especially true for societies where obedience is valued, such as Taiwan and China."



### In brief

- Mani Pande, a research analyst with the Wikimedia Foundation's Global Development Department, announced the final report from the latest Wikipedia Editor Survey. A dump with raw anonymized data <sup>[24]</sup> from the survey was also released by WMF (read the full coverage).
- In an article appearing in the *Communications of the ACM* with the title "Reputation systems for open collaboration", a team of researchers based at UCSC discuss the design of reputational incentives in open collaborative systems and review lessons learned from the development and analysis of two different kinds of reputation tools for Wikipedia (WikiTrust) and for collaborative participation in Google Maps (CrowdSensus).<sup>[25]</sup>
- A paper presented by a Spanish research team at *CISTI 2011* <sup>[26]</sup> presents results from an experiment in using Wikipedia in the classroom and reports on "how the cooperation of Engineering students in a Wikipedia editing project helped to improve their learning and understanding of physics".<sup>[27]</sup>
- Researchers from Karlsruhe Institute of Technology released an analysis and open dataset <sup>[28]</sup> of 33 language corpora extracted from Wikipedia.<sup>[29]</sup>
- A team from the University of Washington and UC Irvine will present a new tool at *WikiSym 2011* for vandalism detection and an analysis of its performance on a corpus of Wikipedia data from the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) <sup>[30]</sup> workshop.<sup>[31]</sup>
- A paper in the *Journal of Oncology Practice*, titled "Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database",<sup>[32]</sup> compared Wikipedia's coverage of ten cancer types with that of Physician Data Query (PDQ) <sup>[33]</sup>, a database of peer-reviewed, patient-oriented summaries about cancer-related subjects which is run by the U.S. National Cancer Institute (NCI). Last year, the main results had already been presented at a conference, announced in a press release <sup>[34]</sup> and summarized in the *Signpost*: "Wikipedia's cancer coverage is reliable and thorough, but not very readable". In addition, the journal

article examines a few other aspects, e.g. that on search engines Google and Bing, "in more than 80% of cases, Wikipedia appeared above PDQ in the results list" for a particular form of cancer.

- A paper published in Springer's *Lecture Notes in Computer Science* presents a new link prediction algorithm for Wikipedia articles and discusses how relevant links to and from new articles can be inferred "from a combination of structural requirements and topical relationships".<sup>[35]</sup>

## References

- [1] K. Nemoto, P. Gloor, and R. Laubacher (2011). Social capital increases efficiency of collaboration among Wikipedia editors. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia – HT '11*, 231. New York, New York, USA: ACM Press. DOI (<http://dx.doi.org/10.1145/1995966.1995997>) • PDF ([http://www.ickn.org/documents/HT2011\\_Nemoto\\_Gloor\\_Laubacher.pdf](http://www.ickn.org/documents/HT2011_Nemoto_Gloor_Laubacher.pdf)) Open access
- [2] <http://www.ht2011.org/>
- [3] A.G. West and I. Lee (2011). What Wikipedia Deletes: Characterizing Dangerous Collaborative Content. In *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*. PDF ([http://www.cis.upenn.edu/~westand/docs/wikisym\\_11\\_revdel\\_final.pdf](http://www.cis.upenn.edu/~westand/docs/wikisym_11_revdel_final.pdf)) Open access
- [4] <http://wikisym.org/ws2011/>
- [5] Alastair Smith (2011). Wikipedia and institutional repositories: an academic symbiosis? In E. Noyons, P. Ngulube and J. Leta (Eds), *Proceedings of the 13th International Conference of the International Society for Scientometrics & Informetrics*, Durban, South Africa, July 4–7, 2011 (pp. 794–800). PDF ([http://www.vuw.ac.nz/staff/alastair\\_smith/publns/SmithAG2011\\_ISSI\\_paper.pdf](http://www.vuw.ac.nz/staff/alastair_smith/publns/SmithAG2011_ISSI_paper.pdf)) Open access
- [6] Dorothea Salo (2008). Innkeeper at the Roach Motel. *Library Trends* 57 (2): 98–12. DOI (<http://dx.doi.org/10.1353/lib.0.0031>) • PDF (<http://digital.library.wisc.edu/1793/22089>) Open access
- [7] <http://www.webcitation.org/61HwRQAF2>
- [8] [http://meta.wikimedia.org/w/index.php?title=File:Expert\\_Participation\\_Survey\\_-\\_Wikimania\\_2011.pdf&page=17](http://meta.wikimedia.org/w/index.php?title=File:Expert_Participation_Survey_-_Wikimania_2011.pdf&page=17)
- [9] <http://siteexplorer.search.yahoo.com/search?p=eprints.otago.ac.nz&bwm=i&bwm=d&bwmf=s>
- [10] <http://www.webcitation.org/61HvQuUT8>
- [11] <http://en.wikipedia.org/w/index.php?title=Special:LinkSearch&target=%2A.eprints.otago.ac.nz>
- [12] <http://www.google.de/search?q=site%3Awikipedia.org+link%3Aeprints.otago.ac.nz>
- [13] <http://linkypedia.inkdroid.org/>
- [14] [https://github.com/jasonpriem/plos\\_altmetrics\\_study/blob/master/crawler/Wikipedia/ArticleStats.php](https://github.com/jasonpriem/plos_altmetrics_study/blob/master/crawler/Wikipedia/ArticleStats.php)
- [15] <http://jhis.sagepub.com/cgi/content/abstract/0165551511416065v1>
- [16] E. Yaari, S. Baruchson-Arbib, and J. Bar-Ilan (2011). Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science* (August 15, 2011). DOI (<http://dx.doi.org/10.1177/0165551511416065>) Closed access
- [17] <http://www.haaretz.com/culture/arts-leisure/is-that-so-1.246189>
- [18] Michaël R. Laurent and Tim J. Vickers (2009). Seeking health information online: does Wikipedia matter? *Journal of the American Medical Informatics Association: JAMIA* 16(4): 471-9 DOI (<http://dx.doi.org/10.1197/jamia.M3059>) • PDF (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705249/pdf/471.S1067502709000802.main.pdf>) Open access
- [19] <http://stats.grok.se/>
- [20] S.T.K. Lam, A. Uduwage, Z. Dong, S. Sen, D.R. Musicant, L. Terveen, and J. Riedl (2011). WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*, 2011. PDF (<http://groupLens.org/system/files/wp-gender-wikisym2011.pdf>) Open access.
- [21] [http://www1.umn.edu/news/news-releases/2011/UR\\_CONTENT\\_350252.html](http://www1.umn.edu/news/news-releases/2011/UR_CONTENT_350252.html)
- [22] Wesley Shu and Yu-Hao Chuang (2011). The Behavior of Wiki Users. *Social Behavior and Personality: An International Journal* 39, no. 6 (October 1, 2011): 851-864. DOI (<http://dx.doi.org/10.2224/sbp.2011.39.6.851>) Closed access
- [23] [http://findarticles.com/p/articles/mi\\_7398/is\\_6\\_39/ai\\_n57926983/](http://findarticles.com/p/articles/mi_7398/is_6_39/ai_n57926983/)
- [24] <http://dumps.wikimedia.org/other/surveys/editorsurvey2011/>
- [25] L. De Alfaro, A. Kulshreshtha, I. Pye, and B. Thomas Adler (2011). Reputation systems for open collaboration. *Communications of the ACM* 54 (8), August 1, 2011: 81. DOI (<http://dx.doi.org/10.1145/1978542.1978560>) • PDF (<http://research.google.com/pubs/archive/36757.pdf>) Open access
- [26] <http://www.aisti.eu/cisti2011/>
- [27] Pilar Mareca, and Vicente Alcober Bosch (2011). Editing the Wikipedia: Its role in science education. In *6th Iberian Conference on Information Systems and Technologies (CISTI)*. HTML ([http://ieeexplore.ieee.org/xpl/freeabs\\_all.jsp?arnumber=5974194](http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5974194)) Closed access.
- [28] <http://km.aifb.kit.edu/sites/corpex/data/>
- [29] Denny Vrandečić, Philipp Sorg, and Rudi Studer (2011). Language resources extracted from Wikipedia. In *Proceedings of the sixth international conference on Knowledge capture - K-CAP '11*, 153. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/1999676.1999703>) • PDF (<http://www.aifb.kit.edu/images/5/5c/Vrandecic11language.pdf>) Open access
- [30] <http://pan.webis.de/>
- [31] Sara Javanmardi, David W McDonald, and Cristina V Lopes (2011). Vandalism Detection in Wikipedia: A High-Performing, Feature-Rich Model and its Reduction Through Lasso. In *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*, 2011. PDF (<http://www.ics.uci.edu/~sjavanma/WikiSym-2011.pdf>) Open access

[32] Malolan S. Rajagopalan, Vineet K. Khanna, Yaacov Leiter, Meghan Stott, Timothy N. Showalter, Adam P. Dicker, and Yaacov R. Lawrence (2011). Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of Oncology Practice* 7(5). **PDF** (<http://www.jopasco.org/site/er/JOP000209.pdf>) • **DOI** (<http://dx.doi.org/10.1200/JOP.2010.000209>) Open access

[33] <http://www.cancer.gov/cancertopics/pdq/cancerdatabase>

[34] <http://www.jeffersonhospital.org/News/2010-june-cancer-information.aspx>

[35] Kelly Itakura, Charles Clarke, Shlomo Geva, Andrew Trotman, and Wei Huang (2011). Topical and Structural Linkage in Wikipedia. In: *Advances in Information Retrieval*, edited by Paul Clough, Colum Foley, Cathal Gurrin, Gareth Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudooh, 6611:460-465. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011 **DOI** ([http://dx.doi.org/10.1007/978-3-642-20161-5\\_45](http://dx.doi.org/10.1007/978-3-642-20161-5_45)) • **PDF** (<http://www.cs.otago.ac.nz/homepages/andrew/2011-1.pdf>) Open access

## Issue 1(3): September 2011

### Top female Wikipedians, reverted newbies, link spam, social influence on admin votes, Wikipedians' weekends, WikiSym previews

**With contributions by:** Tbayer, Daniel Mietchen, DarTar and Jodi.a.schneider

#### What the most active female editors contribute

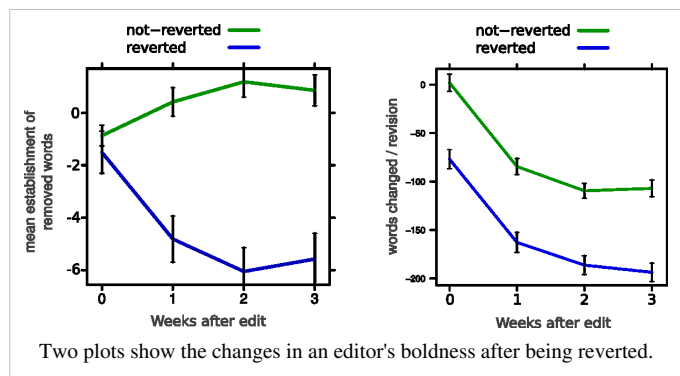
A paper addressing gender imbalance in Wikipedia ("Gender differences in Wikipedia editing") by Judd Antin and collaborators won the "Best Short Paper" award at WikiSym.<sup>[1]</sup> This follows the awarding<sup>[2]</sup> of "best full paper" to another study on the gender gap<sup>[3]</sup> already covered in previous editions of the research newsletter. The study by Antin and collaborators sampled 256,190 users who created a new account on the English Wikipedia between September 2010 and February 2011 and qualitatively coded their contribution by category of wiki work. The results suggest that, whereas in the lower three quartiles by activity level men and women make roughly the same contributions in each category of wiki work, in the top quartile editors behave in a significantly different way. The researchers found that among the top 25% of Wikipedians by activity level:

- only 27% of all revisions are made by women;
- women tend to make larger revisions than men;
- top female editors make significantly larger revisions than men in at least two categories: "adding new content" and "rephrasing existing text"

#### Effects of reverts on wiki work

Another WikiSym 2011 paper by GroupLens researchers, including Summer of Research fellow Aaron Halfaker ("Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work"), reports on the effects of reverts on the quality and quantity of the work of Wikipedia editors, with a specific focus on newbies.<sup>[4]</sup> The study uses a number of key metrics to assess the quality of editor contributions (using *reverts per revision* and *Persistent Word Revisions* or *PWR*, to

measure the survival across revisions of words added by an editor, other than stop-word) and changes in editor activity (using a *controlled activity delta* that calculates an editor's variation of activity across weeks with respect to the week preceding the revert, normalized by the editor's daily rate of activity). The results point at the same time at the important role of reverts as a learning and quality improvement process but also at their negative effects on new contributors. Below are highlights from this study:

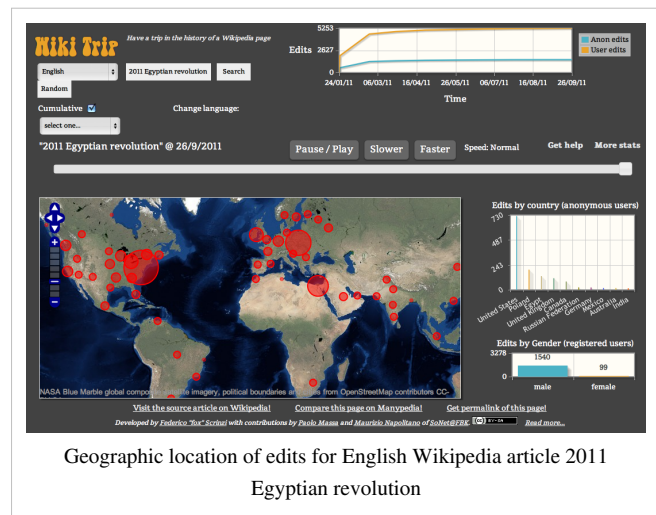


- Compared with their activity prior to a revert, in the first week after a revert reverted editors decrease their activity by 0.1 standard deviations compared with an increase in a control group of non-reverted editors of about 0.3 standard deviations.
- It matters who performs a revert: editors reverted by a registered editor do not recover to the average level of activity for at least one month, whereas editors reverted by anonymous users recover much faster.
- Reverts affect the quality of one's work: reverted editors are less likely to be reverted in the future (particularly in the week after the revert), whereas the probability of being reverted in the control group keeps growing every week. Reverted editors are also less likely to make important changes to an article after being reverted, compared with the control group. However, the productivity of reverted editors in the following weeks increases more rapidly than non-reverted editors.
- Reverts affect newbies more negatively. Experienced editors are less affected by reverts on their average activity while newbies are significantly less likely to continue editing after a revert than experienced editors.

These results are consistent with the findings by Summer of Research fellows on the effects of community interactions with new Wikipedians.

### Further Wikipedia coverage at WikiSym 2011: Social dynamics and global reach

- The two papers on gender gap mentioned above will be presented in a session titled Understanding Wikipedia <sup>[5]</sup>, along with other original works some of which were already reviewed in the research newsletter, such as a study by researchers from the University of Pennsylvania examining revision deletion in the English Wikipedia (see also the summary posted on AcaWiki <sup>[6]</sup>, <sup>[7]</sup>
- A second research session will be devoted to Wikipedia as a global phenomenon <sup>[8]</sup>. It will feature two papers focusing on Wikipedia's coverage of rapidly changing events such as the 2011 Egyptian revolution and the 2011 Tōhoku earthquake and tsunami. The analysis of Wikipedia coverage of the Egyptian revolution, by a team of Italian researchers based in Trento (from the same lab that released the WikiTrip visualization <sup>[9]</sup>, previously covered in the Signpost), is available as a preprint. <sup>[10]</sup>



### Link spam research with controversial genesis but useful results

The "Wiki tools and interfaces <sup>[11]</sup>" session at WikiSym will see the presentation of a paper titled "Autonomous link spam detection in purely collaborative environments <sup>[12]</sup>". According to the five authors from the University of Pennsylvania, link spam is currently "an annoying, but non-pervasive issue", but could become a grave threat to Wikipedia if new spam techniques that were explored by some of them in another paper (see below) become more widespread.

Using the STiki software by one of the authors, which is already widely used as an anti-vandalism tool on the English Wikipedia, the researchers collected mainspace edits adding external links and extracted a corpus of 5,962 link additions classified as either ham or spam, using criteria such as whether the edit had been rolled back (to determine spam), or whether it had been added by a user with rollback rights (to determine ham). From this, the researchers derived numerous features that indicate link spamming behavior, in three areas: On-wiki evidence (including very simple metrics such as the URL's length – spam links tend to be shorter – or that older and more popular articles are more likely to be targeted), properties of the landing page that the link points to (these were

found to be less useful), and classification from third-party sites, including Alexa and Google Safe Browsing. The backlinks data provided by Alexa proved to be most useful for the classifier that the authors went on to construct, and tested in a live implementation in the STiki tool. They conclude that "it is clear this work will benefit the Wikipedia community".

In another paper, presented earlier this month at CEAS '11, five authors from the same university including two of the same researchers examine the possibility of "Link spamming Wikipedia for profit <sup>[13]</sup>". They picture spam detection on Wikipedia as a pipelined process, with the MediaWiki spam blacklist as the first stage (currently containing around 17000 regular expressions), recent changes patrollers (often aided by software tools) as the next – often reacting within seconds after an edit, watchlisters as the third (within minutes to days), and finally review by normal readers as the last stage. Based on a spam/ham corpus constructed as in the other paper, this paper contains some further analysis of the characteristics of link spam destinations and spamming accounts, and of the exposure spammed links receive before they are removed (determined by both the link's lifespan and the popularity of the spammed page). The most sensitive part of the paper then leverages these results to "describe a novel and efficient spam model we estimate can significantly outperform status quo techniques", e.g. by rapidly adding links to exploit the time lag of Wikipedia's spam removal process, or targeting popular pages. In a nod to WP:BEANS, the researchers admit that "there is the possibility that we have introduced previously unknown vectors", but the "Ethical Considerations" section emphasizes that:

"It is in no way this research's intention to facilitate damage to Wikipedia or any wiki host. The vulnerabilities discussed in this section have been disclosed to Wikipedia's parent organization, the Wikimedia Foundation (WMF). Further, the WMF was notified regarding the publication schedule of this document and offered technical assistance."

The authors also point to the implementation of the spam mitigation tool described in the WikiSym article.

However, the paper fails to mention that last year, one of its authors conducted actual, extensive tests of spamming techniques on the English Wikipedia that are very similar to those outlined in the paper. The spam attacks gained the attention of several IT security news websites, and even involved setting up a fake webshop to measure how many Wikipedia readers would have carried out an actual purchase of the penis enlargement pills advertised in the links. The case led to the researcher's temporary ban as a Wikipedia user, later lifted by the arbitration committee, and informed the research guidelines drafted later that year by the Wikimedia Foundation's Research Committee. See *Signpost* coverage: "Large scale vandalism revealed to be 'study' by university researcher" (includes a background interview with the researcher).

## How social ties influence admin votes

A paper by three researchers from the University of the Philippines Diliman<sup>[14]</sup>, presented at the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011) two months ago, examined statistical relations between the voting behavior in requests for adminships (RfAs) and the on-wiki social contacts of participants. The paper includes a brief review of existing literature (in particular two papers which already studied the relation with existing social networks<sup>[15][16]</sup>). Drawing from a January 2008 dump of the English Wikipedia, they analyzed 2,587 elections conducted between 2004 and 2008 (48% of them successful, with 7,231 users voting or running in at least one RfA, and 80% of the final non-neutral votes being supportive), and "1,097,223 instances of communication between 265,155 distinct pairs of users" who had run or voted in an RfA – from user talk page messages, an undirected social graph was generated. Their results concern three areas:

- "Factors that motivate participation": As a first result, the researchers found that the number of a user's contacts who already voted in an RfA, and (more strongly) whether the user had been in contact with the candidate, "contribute positively to the probability of a user's participation in an election. This may be due to the fact that voters are inclined to support candidates with whom they are acquainted with."

- As "Factors that influence voting" (i.e. the support/oppose decision) the authors considered the numbers of "support" and "oppose" votes that a user's contacts have already cast when the user votes, and whether the user had been in contact with the candidate before. All yielded regression coefficients with the expected sign (acquaintance with the candidate weighing positively), and the authors conclude that "we can already explain voting behavior by just examining the immediate neighborhood of a voter", but note that "it is interesting to note that the presence of contacts who have voted negatively weighs more heavily compared with those who voted positively."
- Finally, the paper examined "Influential voters in the social network", by calculating various well-known social network metrics for both the "support" and "oppose" camps in each election ("degree, closeness centrality, betweenness centrality, authority, hub, PageRank, clustering coefficient, and eigenvector centrality", averaged over all voters in each camp, and combined into a weighted difference). Closeness, PageRank, and eigenvector centrality were found to have the largest regression coefficients in predicting the outcome of an RfA, suggesting to the authors "that decisions of influential nodes can affect the outcome of the RfA process. Although it was not studied in this paper, a possible explanation for this result is that influential users may sway other users to vote the same way and this aggregate behavior may have an impact on the result of the election".

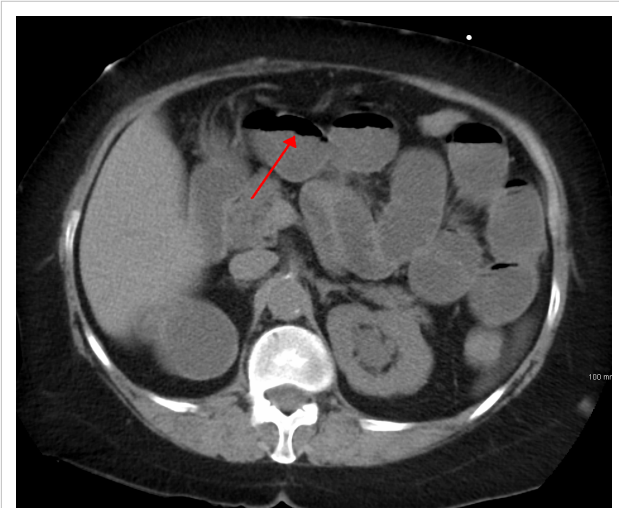
### Wikipedians' weekends in international comparison

A paper titled "Temporal characterization of the requests to Wikipedia" examined how search requests, read accesses and edits on Wikipedia change over time, and relate to those at the entirety of Wikimedia sites (based on squid logs for the whole year of 2009, provided by the Wikimedia Foundation). Among findings are differences between language versions of Wikipedia, such as that the "the number of edits tends to raise in weekends" for the French, Japanese, Dutch and Polish Wikipedia, but not for other languages. Another paper, titled "Circadian patterns of Wikipedia editorial activity: A demographic analysis"<sup>[17]</sup>, similarly analyzed "34 Wikipedias in different languages [trying] to characterize and find the universalities and differences in temporal activity patterns of editors", with the underlying data provided by the German Wikimedia chapter from the toolserver. They found that "in contrast to diurnal [daily] pattern, which is universal to a great extent, weekly activity patterns of WPs show remarkable differences. We could, however, identify two main categories, namely 'weekends' and 'working days' active WPs."<sup>[18]</sup>

### In brief

- **Gender bias in Wikipedia and Britannica:** An article by Joseph Reagle and Lauren Rhue titled "Gender bias in Wikipedia and Britannica" examines gender bias in biographical coverage, comparing the English Wikipedia and the Encyclopedia Britannica.<sup>[19]</sup> The study suggests that "Wikipedia provides better coverage and longer articles, and that it typically has more articles on women than Britannica in absolute terms, but we also find that Wikipedia articles on women are more likely to be missing than are articles on men relative to Britannica". See the accompanying blog post<sup>[20]</sup> with the full datasets used in this study.
- **Wikipedia as a potlatch:** Spanish researcher Felipe Ortega compares Wikipedia to the potlatch, a traditional gift-giving ceremony whose participants gain status based on the generosity of their gifting, in this blog post<sup>[21]</sup> summarizing his new book with Joaquín Rodríguez ( "El Potlatch Digital: Wikipedia y el Triunfo del Procomún y el Conocimiento Compartido" ["The Digital Potlatch: Wikipedia and the Triumph of Commons and Shared Knowledge"], published in Spanish by Ediciones Cátedra.<sup>[22]</sup> Drawing from new qualitative research (interviews with editors of the Spanish Wikipedia) as well as existing quantitative research, the book concludes that recognizing the gifts Wikipedians make, through meritocracy and explicit acknowledgement, helps motivate participation.

**How medical students edit Wikipedia:** A paper published last month by the Kansas Journal of Medicine asked "Are students able and willing to edit Wikipedia to learn components of evidence-based practice?"<sup>[23]</sup> In 2007 and 2008, two groups of senior medical students at the University of Texas Health Science Center at San Antonio participated in an exercise where they were asked "to place succinct summaries of [medical] studies in Wikipedia" (after a four hour introductory course on wikis). In a survey, 91% of them said that the project should be offered again in the next year, and 71% planned to edit Wikipedia again. (The authors caution that this group was self-selected.) The articles were examined two months after their edits, and 46% of the students had their contribution improved in some way, while "the pages edited by 62% of students had additional edits in response to incidental vandalism to the pages, but in no instance was the vandalism done to an edit by a student".



Bowel obstruction was one of the articles edited by the University of Texas students during the course.

- **Ethnography of wikiculture set free:** Joseph Reagle's 2010 book on the cultural dynamics of Wikipedia, *Good Faith Collaboration*, is now freely available<sup>[24]</sup> to read online, having been released under an accommodating Creative Commons licence (CC BY-NC-SA 3.0).<sup>[25]</sup>
- **Provenance for Wikipedia articles:** A (closed access) doctoral dissertation defended at the University of Arizona presents a "domain ontology of provenance for Wikipedia based on the W7 model", building on the notion of the five Ws. The author applies this ontology to extract provenance for Wikipedia articles and to assess their quality, thereby identifying "several collaboration patterns that are preferable or detrimental for data quality".<sup>[26]</sup>
- **Geographies of the World's Knowledge:** as already mentioned in last week's Signpost, the floatingsheep collective<sup>[27]</sup>, in collaboration with the Oxford Internet Institute, released a report titled "Geographies of the World's Knowledge" visualizing the temporal and geographical distribution of Wikipedia articles.<sup>[28]</sup> Drawing from roughly 1.5 million articles in a 2010 database download, the report revealed among other findings that more articles had been written about Antarctica (7,800) than any South American or African nation, that the country with the most internet users (China) accounted for barely 1% of articles, that its biographical articles overwhelmingly geolocate to Western Europe and, from the 18th century on, North America, and that vastly more biographies per year were written for the 20th and particularly the 21st century compared to preceding time periods. The report is released under a Creative Commons BY-NC-ND license.
- **Wikipedia found to have grown until 2007:** A paper by a sociology researcher from the University of York, titled "Measuring the Development of Wikipedia"<sup>[29]</sup>, explores the development of the number of edits and the number of participants on the English Wikipedia from 2002 to 2007 (curiously asserting that "there is only 6 years data"). As first result, the research "reveals that the number of edits and the total number of participants both increased in Wikipedia from 2002 to 2007". The paper's most tangible contribution appears to consist of histograms plotting the number of users with a particular edit count in each of the years 2002 to 2007, which the author finds "are similar with the Pareto distribution in the shape, [and therefore] we assume that the participation situation in Wikipedia is one type of the Pareto Distribution". A large part of the four page paper (available for \$26) is devoted to general explanations of this distribution. It also mentions the need to use a statistical method such as maximum likelihood estimation to confirm the optical impression that the histograms follow the Pareto distribution, but it remains unclear if the author actually carried this out. Also, despite emphasizing several times the importance of determining the changes in the  $k$  parameter over the years (a measure of the inequality



associated with the postulated Pareto distribution) – calling it "vital to model the participation situation in Wikipedia" –, the actual values are never given. The abstract promises "an equation to predict future development trend of Wikipedia", but it remains unclear to this reviewer which equation this refers to.

## References

- [1] Antin, Judd, Raymond Yee, Coye Cheshire, and Oded Nov (2011). Gender Differences in Wikipedia Editing. *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*, 2011. **PDF** (<http://people.ischool.berkeley.edu/~coye/Pubs/Articles/GenderWikiSym2011.pdf>) Open access
- [2] <http://www.wikisym.org/2011/09/21/best-paper-winners-for-wikisym-2011/>
- [3] S.T.K. Lam, A. Uduwage, Z. Dong, S. Sen, D.R. Musicant, L. Terveen, and J. Riedl (2011). WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*, 2011. **PDF** (<http://grouplens.org/system/files/wp-gender-wikisym2011.pdf>) Open access
- [4] Halfaker, Aaron, Aniket Kittur, and John Riedl (2011). Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. *WikiSym '11: Proceedings of the 7th International Symposium on Wikis*. **PDF** (<http://www.grouplens.org/system/files/halfaker11bite.personal.pdf>) Open access
- [5] <http://www.wikisym.org/2011/09/14/session-preview-understanding-wikipedia/>
- [6] [http://acawiki.org/What\\_Wikipedia\\_deletes:\\_Characterizing\\_dangerous\\_collaborative\\_content](http://acawiki.org/What_Wikipedia_deletes:_Characterizing_dangerous_collaborative_content)
- [7] A.G. West and I. Lee (2011). What Wikipedia Deletes: Characterizing Dangerous Collaborative Content. In *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*. **PDF** ([http://www.cis.upenn.edu/~westand/docs/wikisym\\_11\\_revdel\\_final.pdf](http://www.cis.upenn.edu/~westand/docs/wikisym_11_revdel_final.pdf)) Open access
- [8] <http://www.wikisym.org/2011/09/20/session-preview-wikipedia-as-a-global-phenomenon/>
- [9] <http://sonetlab.fbk.eu/wikitrip/>
- [10] Ferron, Michela, and Paolo Massa (2011). Collective memory building in Wikipedia: The case of North African uprisings. *WikiSym 2011: Proceedings of the 7th International Symposium on Wikis*. **PDF** ([http://www.gnuband.org/papers/collective\\_memory\\_building\\_in\\_wikipedia\\_the\\_case\\_of\\_north\\_african\\_uprisings/](http://www.gnuband.org/papers/collective_memory_building_in_wikipedia_the_case_of_north_african_uprisings/)) Open access
- [11] <http://www.wikisym.org/2011/09/19/session-preview-wiki-tools-and-interfaces/>
- [12] [http://www.cis.upenn.edu/~westand/docs/wikisym\\_11\\_spam\\_final.pdf](http://www.cis.upenn.edu/~westand/docs/wikisym_11_spam_final.pdf)
- [13] [http://www.cis.upenn.edu/~westand/docs/ceas\\_11\\_wiki\\_spam\\_final.pdf](http://www.cis.upenn.edu/~westand/docs/ceas_11_wiki_spam_final.pdf)
- [14] Cabunducan, Gerard, Ralph Castillo, and John Boaz Lee (2011). Voting behavior analysis in the election of Wikipedia admins. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*, 545–547. IEEE **DOI** (<http://dx.doi.org/10.1109/ASONAM.2011.42>) Closed access
- [15] J. Leskovec, D. Huttenlocher, J. Kleinberg (2010) Predicting positive and negative links in online social networks. *ACM WWW International conference on World Wide Web (WWW '10)*, 2010. video ([http://videolectures.net/www2010\\_leskovec\\_ppn/](http://videolectures.net/www2010_leskovec_ppn/)) **PDF** (<http://cs.stanford.edu/people/jure/pubs/signs-www10.pdf>) Open access
- [16] J. Leskovec, D. Huttenlocher, J. Kleinberg (2010) Governance in Social Media: A case study of the Wikipedia promotion process. In: *AAAI International Conference on Weblogs and Social Media (ICWSM '10)*. video ([http://videolectures.net/icws2010\\_leskovec\\_gsm/](http://videolectures.net/icws2010_leskovec_gsm/)) **PDF** (<http://cs.stanford.edu/people/jure/pubs/voting-icws2010.pdf>) Open access
- [17] Yasseri, Taha, Sumi, Róbert, Kerétsz, János (2011). Circadian patterns of Wikipedia editorial activity: A demographic analysis, *ArXiv* (September 8, 2011). **PDF** (<http://arxiv.org/abs/1109.1746>) Open access
- [18] Reinoso, Antonio J., Jesus M. Gonzalez-Barahona, Rocio Muñoz-Mansilla, and Israel Herraiz (2011). Temporal characterization of the requests to Wikipedia. In *Proceedings of the 5th International Workshop on New Challenges in Distributed Information Filtering and Retrieval (DART 2011)*. ETSI Caminos, Canales y Puertos (UPM), September 13, 2011. **PDF** (<http://oa.upm.es/8836/1/temporalCharac.pdf>) Open access
- [19] Reagle, Joseph, and Lauren Rhue (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication* 5 (2011): 1138–1158. **PDF** (<http://ijoc.org/ojs/index.php/ijoc/article/view/777>) Open access
- [20] <http://reagle.org/joseph/blog/social/wikipedia/gender-bias-in-wp-eb>
- [21] <http://blog.felipeortega.net/2011/09/20/the-digital-potlatch/>
- [22] José Felipe Ortega and Joaquín Rodríguez López (2011). *El potlatch digital. Wikipedia y el triunfo del procomún y el conocimiento compartido*, Catedra, September 2011. **HTML** (<http://www.catedra.com/cgi/general/newFichaProducto.pl?obrcod=2885118>) Closed access
- [23] Badgett, Robert G, and Mary Moore (2011). Are students able and willing to edit Wikipedia to learn components of evidence-based practice? *Kansas Journal of Medicine* 4(3), August 30, 2011. **PDF** (<http://hdl.handle.net/2271/976>) Open access
- [24] <http://reagle.org/joseph/2010/gfc/>
- [25] Reagle, Joseph M. (2010). *Good Faith Collaboration: The Culture of Wikipedia*. The MIT Press, 2010. **HTML** (<http://reagle.org/joseph/2010/gfc/>) Open access
- [26] Liu, J. (2011). *W7 model of provenance and its use in the context of Wikipedia*. PhD dissertation, The University of Arizona, 2011. **PDF** (<http://gradworks.umi.com/34/60/3460890.html>) Closed access
- [27] <http://www.floatingsheep.org/2011/09/two-of-floating-sheep-collective-have.html>

[28] Graham, M., Hale, S. A. and Stephens, M. (2011) *Geographies of the World's Knowledge*. Ed. Flick, C. M., London, Convoco! Edition.

PDF ([http://www.oii.ox.ac.uk/publications/convoco\\_geographies\\_en.pdf](http://www.oii.ox.ac.uk/publications/convoco_geographies_en.pdf)) Open access

[29] He, Zeyi (2011). Measuring the Development of Wikipedia. In *2011 International Conference on Internet Technology and Applications*,

IEEE DOI (<http://dx.doi.org/10.1109/ITAP.2011.6006393>) Closed access

## Issue 1(4): October 2011

### WikiSym; predicting editor survival; drug information found lacking; RfAs and trust; Wikipedia's search engine ranking justified

With contributions by: Boghog, Jodi.a.schneider, Drdee, DarTar, Phoebe and Tbayer

#### Wiki research beyond the English Wikipedia at WikiSym

WikiSym 2011, the "7th international symposium on wikis and open collaboration", took place from October 3-5 at the Microsoft Research Campus in Silicon Valley (Mountain View, California). Although the conference's scope has broadened to include the study of open online collaborations that are not wiki-based, Wikipedia-related research still took up a large part of the schedule <sup>[1]</sup>. Several of the conference papers have already been reviewed in the September and August issues of this research overview, and the rest of the proceedings <sup>[2]</sup> have since become available online.



The workshop " **WikiLit: Collecting the Wiki and Wikipedia Literature** <sup>[3],[4]</sup>, led by Phoebe Ayers and Reid Priedhorsky, explored the daunting task of collecting the scholarly literature pertaining to Wikipedia and wikis generally. Research about wikis can be difficult to find, since there are papers published in many fields (from sociology to computer science) and in many formats, from published articles to on-wiki community documents. There have been several attempts over the years to collect the wiki and Wikipedia literature, including on Wikipedia itself, but all such projects have suffered from not keeping up to date with the sheer volume of research that is published every year. While the workshop did not reach consensus on what platform to proceed with to build a sustainable system, there was agreement that this is an important topic for the research and practitioner community, and the group developed a list of requirements <sup>[3]</sup> that such a system should have. The workshop followed and extended discussions on the wiki-research-1 <sup>[5]</sup> mailing list earlier this year on the topic.

In a panel titled " **Apples to Oranges?: Comparing across studies of open collaboration/peer production** <sup>[6],[7]</sup>, six US-based scholars reviewed the state of this field of research. Among the takeaways were a call to study failed collaboration projects more often instead of focusing research on successful "anomalies" like Wikipedia, and – especially in the case of Wikipedia – to broaden research to non-English projects.

Another workshop, titled " **Lessons from the classroom: successful techniques for teaching wikis using Wikipedia** <sup>[8],[9]</sup> was a retrospective on the Wikimedia Foundation's Public Policy Initiative.

Among the conference papers not mentioned before in this newsletter are:

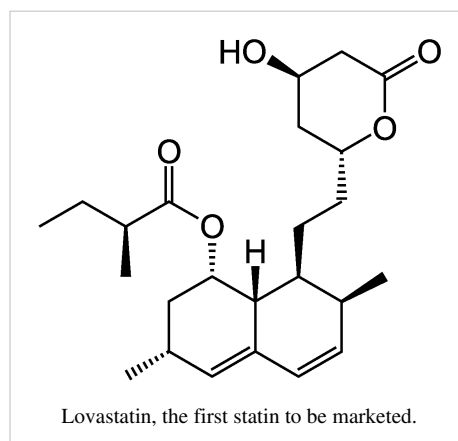
- " **Mentoring in Wikipedia: a clash of cultures** " <sup>[10]</sup>, a paper which "draw[s] insights from the offline mentoring literature to analyze mentoring practices in Wikipedia and how they influence editor behaviors. Our quantitative analysis of the Adopt-a-user program shows mixed success of the program".
- " **Vandalism Detection in Wikipedia: A High-Performing, Feature-Rich Model and its Reduction** " <sup>[11]</sup> – arguing that on Wikipedia "human vigilance is not enough to combat vandalism, and tools that detect possible vandalism and poor-quality contributions become a necessity", the authors present a vandalism classifier constructed using machine learning techniques.

Wikipedia-related posters included

- "A scourge to the pillar of neutrality: a WikiProject fighting systemic bias"<sup>[12]</sup> presenting preliminary findings from an ongoing survey and interviews among members of the WikiProject Countering systemic bias.
- Another poster presentation planned to analyze the contributions of the members of this WikiProject to see what kind of systemic bias they might exhibit themselves ("Places on the map and in the cloud: representations of locality and geography in Wikipedia"<sup>[13]</sup>).
- "Participation in Wikipedia's article deletion processes"<sup>[14]</sup> found that "the deletion process is heavily frequented by a relatively small number of longstanding users" and that "the vast majority of [speedily] deleted articles are not spam, vandalism, or 'patent nonsense', but rather articles which could be considered encyclopedic, but do not fit the project's standards".
- "Exploring underproduction in Wikipedia"<sup>[15]</sup> examined "two key circumstances in which collective production can fail to respond to social need: when goods fail to attain high quality despite (1) high demand or (2) explicit designation by producers as highly important".

### Quality of drug information in Wikipedia

A study entitled "Accuracy and completeness of drug information in Wikipedia: an assessment"<sup>[16]</sup> in this month's issue of the *Journal of the Medical Library Association* of five widely prescribed statins found that while these Wikipedia drug articles are generally accurate, they are incomplete and inconsistent. The study's authors conclude:



Because the entries on the five most commonly prescribed statins lacked important information, the authors recommend that consumers should seek other sources and not rely solely on Wikipedia.

The main criticism by the study is that most of the articles lacked sufficient information on adverse effects, contraindications, and drug interactions and this lack of information might harm the consumer. These criticisms echo earlier ones (two similar studies reported in the *Signpost*: "Pharmacological study criticizes reliability of Wikipedia articles about the top 20 drugs", "Wikipedia drug coverage compared to Medscape, found wanting"). However the authors did note the benefit of Wikipedia hypertext links to additional information that most other web sources on drug information lack and in addition noted that all the Wikipedia articles contained references to peer reviewed journals and other reliable sources. Hence overall, the latest study is somewhat more positive than the earlier two.

## Predicting editor survival: The winners of the Wikipedia Participation Challenge

The Wikimedia Foundation announced the winner of the Wikipedia Participation Challenge. The data competition, organized in partnership with Kaggle and the 2011 IEEE International Conference on Data Mining [17], asked data scientists to use Wikipedia editor data and develop an algorithm to predict the number of future edits, and in particular one that correctly predicts who will stop editing and who will continue to edit (see the call for submissions [18]). The response was overwhelming, with 96 participating teams, comprising in total 193 people who jointly submitted 1029 entries (listed in the competition's leaderboard [19]).

The brothers Ben and Fridolin Roth (from team prognozIt) developed the winning algorithm. They developed a linear regression model using Python and GNU Octave. The algorithm used 13 features (2 based on reverts and 11 based on past editing behavior) to predict future editing activity. Both the source code [20] and a description of the algorithm are available. Unfortunately, because it relied on patterns in the training dataset that would not be present in the actual one, the model's ongoing use is severely restricted.

Second place went to Keith Herring [21]. Submitting only 3 entries, he developed a highly accurate model, using random forests, and utilizing a total of 206 features. His model shows that a randomly selected Wikipedia editor who has been active in the past year has approximately an 85 percent probability of becoming inactive (no new edits) in the following 5 months. The most informative features captured both the edit timing and volume of an editor's activity.

The challenge also announced two Honourable Mentions for participants who only used open source software. The first Honourable Mention went to Dell Zang [22] (team zeditor) who used a machine learning technique called gradient boosting. His model mainly uses recent past editor activity. The second Honourable Mention went to Roopesh Ranjan and Kalpit Desai (team Aardvarks). Using Python and R, they too developed a random forest model. Their model used 113 features, mainly based on the number of reverts and past editor activity (see its full description).

All the documentation and source code has been made available on the main entry page for the WikiChallenge.

## What it takes to become an admin: Insights from the Polish Wikipedia

A team of researchers based at the Polish Japanese Institute of Information Technology [23] (PJIT) published a study presented at *SocInfo 2011* looking at Requests for Adminship (RfA) discussions in the Polish Wikipedia. [24] The paper presents a number of statistics about adminship in the Polish Wikipedia since the RfA procedure was formalized (2005), including the rejection rate of candidates across different rounds, the number of candidates and votes over the years and the distribution of tenure and experience of candidates for adminship. The results indicate that it was far more complicated to obtain admin status in 2010 than it was in previous years, and that tenure required to be a successful RfA candidate has soared dramatically: "the mean number of days since registration to receiving adminship is nearly five times larger than it was five years before".

The remainder of the paper studies RfA discussions by comparing the social network of participants based on their endorsement (vote-for) or rejection (vote-against) of a given candidate with an implicit social network derived from



three different types of relations between contributors (trust, criticism and acquaintance). The goal is to measure to what extent these different kinds of relations can predict voting behavior in the context of RfA discussions. The findings suggest that "trust" and "acquaintance" (measured respectively as the amount of edits by an editor in the vicinity of those by the other editor and as the amount of discussions between two contributors) are significantly higher in votes-for than in votes-against. Conversely, "criticism" (measured as the number of edits made by one author and reverted by another editor) is significantly higher in votes-against than in votes-for.

This study complements research on the influence of social ties on adminship discussions reviewed in the past edition of the research newsletter.

### **High search engine rankings of Wikipedia articles found to be justified by quality**

An article titled "Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality?", by two professors for information research from the Hamburg University of Applied Sciences (which appeared earlier this year in the *Journal of the American Society for Information Science and Technology* and is now available as open access, also in form of a recent arxiv preprint<sup>[25]</sup>) addresses "the fiercely discussed question of whether the ranking of Wikipedia articles in search engines is justified by the quality of the articles". The authors recall an earlier paper coauthored by one of them<sup>[26]</sup> that had found Wikipedia to be "by far the most popular" host in search engine results pages (in the US): In "1000 queries, Yahoo showed the most Wikipedia results within the top 10 lists (446), followed by MSN/Live (387), Google (328), and Ask.com (255)". They then set out to investigate "whether this heavy placement is justified from the user's perspective". First, they re-purposed the results of a 2008 paper of the first author,<sup>[27]</sup> where students had been asked to judge the relevance of search engine results for 40 queries collected in 2007, restricting them to the search results that consisted of Wikipedia articles – all of them from the German version. They found that "Wikipedia results are judged much better than the average results at the same ranking position" by the jurors, and that

“The data indicates that contrary to the assumption that Wikipedia articles show up too often in the search engines' results, the search engines could even think of improving their results through providing more Wikipedia results in the top positions.”

In order to conduct a more thorough investigation (the 2008 assessments having only focused on the criterion of relevance), the present paper sets out to develop a set of quality criteria for the evaluation of Wikipedia articles by human jurors. It first gives an overview of existing literature about the information quality of Wikipedia, and of encyclopedias in general, identifying four main criteria that several pre-2002 works about the quality of reference works agreed on. Interestingly, "accuracy" was not among them, an omission explained by the authors by the difficulty of fact-checking an entire encyclopedia. From this, the authors derive a set of 14 evaluation criteria, incorporating both the general criteria from the literature about reference works and internal Wikipedia criteria such as the status of being a featured/good article, the verifiability of the content and the absence of original research. These were then applied by the jurors (two last year undergraduate students with experience in similar coding tasks) to 43 German Wikipedia articles that had appeared in the 2007 queries, in their state at that time. While "the evaluated Wikipedia articles achieve a good score overall", there were "noticeable differences in quality among the examples in the sample" (the paper contains interesting discussions of several strengths and weaknesses according to the criteria set, e.g. the conjecture that the low score on "descriptive, inspiring/interesting" writing could be attributed to "the German academic style. A random comparison with the English version of individual articles seems to support this interpretation").

The authors conclude:

“In general, our study could confirm that the ranking of Wikipedia articles in search engines is justified by a satisfactory overall quality of the articles. ... In answer to research question 4b, 4c ('Is the ranking appropriate? Are good entries ranked high enough?'), we can say that the rankings in search engines are at least appropriate.”

Both the search engine ranking data and the evaluated Wikipedia article revisions are somewhat dated, referring to January 2007 (the authors themselves note that it "could well be that in the meantime search engines reacted to that fact [the potential of improving results by ranking Wikipedia higher] and further boosted Wikipedia results", and also that regarding the German Wikipedia, the search engine results did not take into account possible effects of the introduction of stable versions in 2008).

### Attempts to predict the outcome of AfD discussions from an article's edit history

A master's thesis defended by Ashish Kumar Ashok, a student in computing at Kansas State University, describes machine learning methods to determine how the final outcome of an Article for Deletion (AfD) discussion is affected by the editing history of the article.<sup>[28]</sup> The thesis considers features such as the structure of the graph of revisions of an article (based on text changed, added or removed), the number of edits of the article, the number of disjoint edits (according to some contiguity definition), as well as properties of the corresponding AfD, such as the number of */votes* and the total length of words used by participants in AfD who expressed their preference to *keep*, *merge* or *delete* the article. Different types of classifiers based on the above features are applied to a small sample of 64 AfD discussions from the 1 August 2011 deletion log. The results of the analysis indicate that the performance of the classifiers does not significantly improve by considering any of the above features in addition to the sheer number of */votes*, which limits the scope and applicability of the methods explored in this work to predict the outcome of AfD discussions. The author suggests that datasets larger than the sample considered in this study should be obtained in order to assess the validity of these methods.

### In brief

- **Why did Wikipedia succeed while others failed?:** In a presentation on October 11 at the Berkman Center for Internet and Society ("Almost Wikipedia: What Eight Collaborative Encyclopedia Projects Reveal About Mechanisms of Collective Action<sup>[29]</sup>", with video), MIT researcher and Wikimedia Foundation advisory board member Benjamin Mako Hill presented preliminary results of his research comparing Wikipedia and seven other Internet encyclopedia projects or proposals that did not take off, based on interviews with the projects' founders as well as examinations of their archives. The event was summarised<sup>[30]</sup> for the Nieman Journalism Lab (reprinted<sup>[31]</sup> in *Business Insider*), and in the *Signpost*: "The little online encyclopaedia that could". Hill later gave a shorter (ca. 12min) talk about the same topic at the "Digital commons" forum (see below): video<sup>[32]</sup>, slides<sup>[33]</sup>.
- **"Digital Commons" conference:** On October 29-30, the "Building Digital Commons<sup>[34]</sup>" conference took place in Barcelona, organized by Catalan Wikimedians and Wikimedia Research Committee member Mayo Fuster Morell, and supported by the Wikimedia Foundation. The program<sup>[35]</sup> featured several presentations about Wikipedia research; further online documentation is expected to become available later.
- **Placement of categories examined:** A paper from two computer science researchers based at the Katholieke Universiteit Leuven examines the order in which categories are placed on a Wikipedia article and reports on connections between a category's position in this list and "its persistence within the article, age, popularity, size, and descriptiveness".<sup>[36]</sup> The order in which categories are added is not determined by any explicit rule. However, the research found, older, more persistent and more exclusive categories are consistently placed in lower positions. Categories appearing at lower positions also tend to do so across all the articles they contain and they include articles that are more similar to each other in terms of category overlap.
- **Visualizing semantic data:** A team from the UCSB Department of Computer Science recently presented<sup>[37]</sup> WiGiPedia<sup>[38]</sup>, a tool visualizing rich semantic data about Wikipedia articles, designed to "inform the user of interesting contextual information pertaining to the current article, and to provide a simple way to introduce and/or repair semantic relations between wiki articles". The tool builds on structured data represented via templates, categories and infoboxes and queried via DBpedia. By supporting collaborative editing of rich semantic data and one-click semantic updates of Wikipedia articles, the tool aims to bridge the gap between Wikipedia and DBpedia. The source code of the tool doesn't appear to be publicly released.

- **Wikipedia literature review:** Owen S. Martin posted to arXiv a 28-page Wikipedia literature review<sup>[39]</sup> towards his Ph.D. in statistics.<sup>[40]</sup> About half the paper gives an overview of Wikipedia's database structure; the remainder reviews about 30 recent papers from the perspective of assessing their quality, trust, semantic extraction, governance, economic implications and epistemological implications.
- **Vandalism detection contest:** An "Overview of the 2nd International Competition on Wikipedia Vandalism Detection" has been published.<sup>[41]</sup>
- **Matching Wikipedia articles to Geonames entries:** A four-page paper by two researchers from Hokkaido University<sup>[42]</sup> explored the problem of "merging Wikipedia's Geo-entities and GeoNames" to form a larger geographical database. This is already being done by the YAGO (Yet Another Great Ontology) database, but the paper uses additional data beyond the article name, such as categories and disambiguation pages on Wikipedia, in order to identify further matching pairs missed in YAGO (and in the process found several errors in GeoNames).
- **Attempt to examine evolution of key activities in Wikipedia:** A paper titled "Governing Complex Social Production in the Internet: The Emergence of a Collective Capability in Wikipedia"<sup>[43]</sup> (presented last month at the "Decade in Internet Time" symposium at the Oxford Internet Institute) undertakes "an exploratory theoretical analysis to clarify the structure and mechanisms driving the endogenous change of [Wikipedia]", using the framework of capability theory to construct six hypotheses such as "the membership in group(s) of contributors that take up governance tasks varies less than in those revolving on content production". These are then tested empirically by applying a clustering algorithm to monthly snapshots of the English Wikipedia (until 2009) "to identify distinct groupings of contributors at each month". However, the clustering algorithm leaves out a group of users "that covers all the observed domains of activity" and "despite its relatively small share of overall contributor population ... provides the majority of the work", which leads the authors to dub it "the core editors of Wikipedia".

## References

- [1] <http://www.wikisym.org/ws2011/program:schedule>
- [2] <http://www.wikisym.org/ws2011/program:proceedings>
- [3] <http://www.wikisym.org/ws2011/workshop:wikilit>
- [4] Ayers, Phoebe, and Reid Priedhorsky (2011). WikiLit: Collecting the wiki and Wikipedia literature. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 229. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038612>) • PDF ([http://wikisym.org/ws2011/\\_media/proceedings:p229-ayers.pdf](http://wikisym.org/ws2011/_media/proceedings:p229-ayers.pdf)) Open access
- [5] <https://lists.wikimedia.org/mailman/listinfo/wiki-research-l>
- [6] <http://wikisym.org/ws2011/proceedings:p227-yew>
- [7] Antin, Judd, Ed H. Chi, James Howison, Sharoda Paul, Aaron Shaw, and Jude Yew (2011). Apples to oranges? Comparing across studies of open collaboration/peer production. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 227. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038610>) • PDF ([http://wikisym.org/ws2011/\\_media/proceedings:p227-yew.pdf](http://wikisym.org/ws2011/_media/proceedings:p227-yew.pdf)) Open access
- [8] <http://wikisym.org/ws2011/proceedings:p231-schulenburg>
- [9] Schulenburg, Frank, LiAnna Davis, and Max Klein (2011) Lessons from the classroom. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 231. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038613>) • PDF ([http://www.wikisym.org/ws2011/\\_media/proceedings:p231-schulenburg.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p231-schulenburg.pdf)) Open access
- [10] Musicant, David R., Yuqing Ren, James A. Johnson, and John Riedl (2011). Mentoring in Wikipedia: a clash of cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 173. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038586>) • PDF ([http://www.wikisym.org/ws2011/\\_media/proceedings:p173-musicant.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p173-musicant.pdf)) Open access
- [11] Javanmardi, Sara, David W. McDonald, and Cristina V. Lopes (2011). Vandalism detection in Wikipedia. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 82. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038573>) • PDF ([http://www.wikisym.org/ws2011/\\_media/proceedings:p82-javanmardi.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p82-javanmardi.pdf)) Open access
- [12] Livingstone, Randall M (2011). A scourge to the pillar of neutrality. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 209. New York, New York, USA: ACM Press, 2011. DOI (<http://dx.doi.org/10.1145/2038558.2038597>) • PDF ([http://www.wikisym.org/ws2011/\\_media/proceedings:p209-livingstone.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p209-livingstone.pdf)) Open access
- [13] Livingstone, Randall M. (2011) Places on the map and in the cloud: representations of locality and geography in Wikipedia. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 211. New York, New York, USA: ACM Press, 2011.

- DOI** (<http://dx.doi.org/10.1145/2038558.2038598>) **PDF** ([http://www.wikisym.org/ws2011/\\_media/proceedings:p211-livingstone.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p211-livingstone.pdf)) Open access
- [14] Geiger, R. Stuart, and Heather Ford (2011). Participation in Wikipedia's article deletion processes. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 201. New York, New York, USA: ACM Press, 2011. **DOI** (<http://dx.doi.org/10.1145/2038558.2038593>) • **PDF** ([http://www.wikisym.org/ws2011/\\_media/proceedings:p201-geiger.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p201-geiger.pdf)) Open access
- [15] Gorbatai, Andreea D. (2011) Exploring underproduction in Wikipedia. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, 205. New York, New York, USA: ACM Press, 2011. **DOI** (<http://dx.doi.org/10.1145/2038558.2038595>) • **PDF** ([http://www.wikisym.org/ws2011/\\_media/proceedings:p205-gorbatai.pdf](http://www.wikisym.org/ws2011/_media/proceedings:p205-gorbatai.pdf))
- [16] Kupferberg, Natalie, and Bridget McCrate Protus (2011) Accuracy and completeness of drug information in Wikipedia: an assessment. *Journal of the Medical Library Association* 99(4): 310-3. **DOI** (<http://dx.doi.org/10.3163/1536-5050.99.4.010>) • **HTML** (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3193353/>) Closed access
- [17] <http://www.eecs.wsu.edu/~holder/icdm2011contest/>
- [18] <http://blog.wikimedia.org/2011/06/28/data-competition-announcing-the-wikipedia-participation-challenge/>
- [19] <http://www.kaggle.com/c/wikichallenge/Leaderboard>
- [20] <http://dumps.wikimedia.org/other/wikichallenge/>
- [21] <http://blog.kaggle.com/2011/10/06/like-popping-bubble-wrap/>
- [22] <http://blog.kaggle.com/2011/10/26/long-live-wikipedia-dell-zhang/>
- [23] <http://www.pjwstk.edu.pl/en/>
- [24] Turek, Piotr, Justyna Spychała, Adam Wierzbicki, and Piotr Gackowski (2011) Social Mechanism of Granting Trust Basing on Polish Wikipedia Requests for Adminship. In: *Social Informatics 2011*. Lecture Notes in Computer Science, 6984:212-225. **DOI** ([http://dx.doi.org/10.1007/978-3-642-24704-0\\_25](http://dx.doi.org/10.1007/978-3-642-24704-0_25)) Closed access
- [25] Lewandowski, Dirk, and Ulrike Spree (2011) Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? *Journal of the American Society for Information Science* 62(1): 117-132. **DOI** (<http://dx.doi.org/10.1002/asi.21423>) Open access • [arxiv.org](http://arxiv.org) **PDF** (<http://arxiv.org/abs/1109.0916>) Open access
- [26] Höchstötter, Nadine, and Dirk Lewandowski (2009). What users see – Structures in search engine results pages. *Information Sciences* 179 (12): 1796-1812 **DOI** (<http://dx.doi.org/10.1016/j.ins.2009.01.028>) • **PDF** (<http://hdl.handle.net/10760/16081>) Open access
- [27] Lewandowski, Dirk (2008). The retrieval effectiveness of Web search engines: Considering results descriptions. *Journal of Documentation* 64(6), 915-937 **PDF** (<http://hdl.handle.net/10760/11258>) Open access
- [28] Ashok, Ashish Kumar (2011). *Predictive data mining in a collaborative editing system: the Wikipedia articles for deletion process*. **HTML** (<http://hdl.handle.net/2097/12026>) Open access
- [29] <http://cyber.law.harvard.edu/events/luncheon/2011/10/makohill>
- [30] <http://www.niemanlab.org/2011/10/the-contribution-conundrum-why-did-wikipedia-succeed-while-other-encyclopedias-failed/>
- [31] [http://articles.businessinsider.com/2011-10-13/tech/30274391\\_1\\_wikipedia-encyclopedias-yochai-benkler](http://articles.businessinsider.com/2011-10-13/tech/30274391_1_wikipedia-encyclopedias-yochai-benkler)
- [32] [http://epicenter.media.mit.edu/~mako/digcom/hill-digcom-talking\\_head.ogv](http://epicenter.media.mit.edu/~mako/digcom/hill-digcom-talking_head.ogv)
- [33] <http://epicenter.media.mit.edu/~mako/digcom/hill-digcom-slides.ogv>
- [34] [http://wiki.digital-commons.net/Main\\_Page](http://wiki.digital-commons.net/Main_Page)
- [35] <http://www.digital-commons.net/program/>
- [36] Gyllstrom, Karl, and Marie-Francine Moens (2011) Examining the "Leftness" Property of Wikipedia Categories. In: CIKM '11. **PDF** (<https://lirias.kuleuven.be/bitstream/123456789/318601/1/GyllstromCIKM2011.pdf>) Open access
- [37] Bostandjiev, Svetlin, John O'Donovan, Brynjar Gretarsson, Christopher Hall, and Tobias Hollerer (2011) WiGiPedia: Visual Editing of Semantic Data in Wikipedia. In: *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2011)*, **PDF** (<http://bostandjiev.com/Content/publications/bostandjiev.pdf>) Open access
- [38] <http://www.wikipedia-online.com>
- [39] <http://arxiv.org/abs/1110.5863>
- [40] Martin, Owen S (2011) A Wikipedia Literature Review. *ArXiv*, October 17, 2011. **PDF** (<http://arxiv.org/pdf/1110.5863v1>) Open access
- [41] Potthast, Martin, and Teresa Holfeld (2011) Overview of the 2nd International Competition on Wikipedia Vandalism Detection. In: PAN 2011. **PDF** ([http://www.uni-weimar.de/medien/webis/publications/papers/stein\\_2011u.pdf](http://www.uni-weimar.de/medien/webis/publications/papers/stein_2011u.pdf)) Open access
- [42] Yiqi Liu, and Masaharu Yoshioka (2011) Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames. **PDF** (<http://sigsw.org/papers/SIG-SWO-A1102/SIG-SWO-A1102-03.pdf>) Open access
- [43] Aaltonen, Aleksí, and Giovan Francesco Lanzara (2011) Governing Complex Social Production in the Internet: The Emergence of a Collective Capability in Wikipedia. In *Decade in Internet Time symposium*. **HTML** (<http://ssrn.com/paper=1926138>) Open access



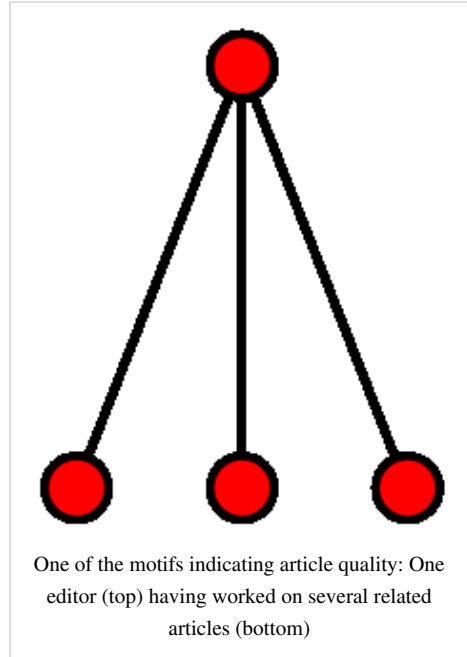
## Issue 1(5): November 2011

### Quantifying quality collaboration patterns, systemic bias, POV pushing, the impact of news events, and editors' reputation

With contributions by: Tbayer, Hfordsa, DarTar and Romanesco

#### Collaboration pattern analysis: Editor experience more important than "many eyes"

A paper titled "Characterizing Wikipedia Pages using Edit Network Motif Profiles"<sup>[1]</sup> by three researchers from University College Dublin indicates that the quality of a Wikipedia article can be predicted from characteristics of its "edit network" – a graph derived from the collaboration of Wikipedians in that area. Network motifs are small graphs which occur particularly frequently as sub-graphs of networks of a certain kind, and can be regarded as its building blocks in some sense. (The concept is popular in bioinformatics, where it is applied to gene regulatory networks.) In this paper, the authors use graphs with at most five nodes consisting of users and articles, which are connected by an edge if the user has edited the article – giving 17 possible "Wikipedia network motifs". (Anonymous users are disregarded.) For a Wikipedia article, the researchers form an "ego network" consisting of that article, articles which link to it (and have been edited by at least one of the users who edited the core article), and the users who edited them. For a sample of around 2000 articles from the History and United States categories, the frequencies of the 17 "Wikipedia network motifs" in those article's "ego networks" were calculated.



Using machine learning techniques, the researchers are able to discern with some certainty articles of basic quality (defined as having been assessed as Start class by Wikipedians) from those of good quality (defined as Featured or B class), solely based on this set of motif frequencies in the article's edit network. Looking at the impact of each of the 17 types separately, they found that "all network motifs have some potential to discriminate between good and basic Wikipedia articles" in the sample, but that among the four best predicting motifs, three are "stars with editors at their centre":

*"This is interesting because it shows that many eyes is not really the defining characteristic of quality; instead experience is important – the editors should have worked on many other articles."*

Another section of the paper constructs spatializations of the sample (i.e. a two-dimensional mapping where articles with similar motif frequency are close to each other). For the history articles sample, this visualization clearly separated B class and Start class articles, but Featured articles are "more spread out", with two clusters on opposite sides of the diagram. The researchers made the interesting discovery that this seems related to the assessed importance of the articles:

*"It transpires that the Featured Articles on the left are inclined to be low or mid importance compared to high or top importance articles on the right. This niche characteristic is emphasized by the fact that these articles are inclined not to have been featured on the Wikipedia main page. We conclude from this that, at least in edit network terms, some low importance Featured Articles look like more ordinary articles. ... It seems that articles on niche topics can reach Featured Article status without a huge amount of collaboration."*

## Systemic bias quantified for twenty language Wikipedias

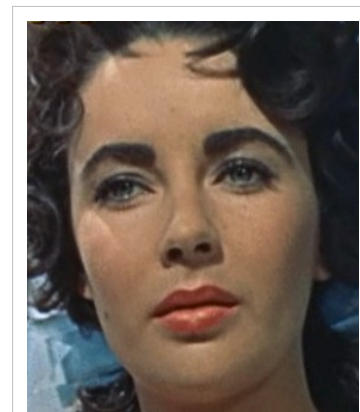
A paper titled "Cultural Configuration of Wikipedia: Measuring Autoreferentiality in Different Languages"<sup>[2]</sup> by two researchers from the Universitat Politècnica de Catalunya, published in the proceedings of the "Recent Advances in Natural Language Processing" conference and apparently based on the first author's masters thesis<sup>[3]</sup> attempts to test the hypothesis that "contributing for the visibility of the own national or language related content" is among the motivations to participate in Wikipedia. According to the authors, "some informal surveys in Catalan WP association 'Amical Viquipèdia' showed how the national topics were a focus of interest for writing and conflict". They propose the concept of "autoreferentiality" "to describe the interest of a culture on itself, which in WP translates to the interest of editors for their own local content in a WP language edition", and set out to measure it by various quantitative features, which are first defined on the article level and then tested on a selection of articles that are assumed to be "local content", using the Java-based WikAPIdia<sup>[4]</sup> tool. (This set is formed by starting with a few keywords clearly pertaining to the local language, and then including articles which share categories – as examples from their own language, the authors list "catalunya", "català", and also "valencia" or "mallorquí" as start words, which "would retrieve titles in articles and categories like "escriptors de catalunya" or "dret català", referring to writers and law".) Among the tested quantitative features are:

- "Isolation", based on the number of interwiki links
- "Effort", based on the size of the article and the number of internal links it contains
- "Prominence", based on the number of incoming wikilinks, the number of categories where the article is a member, and its PageRank
- "Edition", a measure of how diverse the authorship of an article is, specified as the smallest number of editors who together contribute 80% of the page's edits (assuming to be lower for local content because it is edited by "highly motivated users")

The paper applies the eventual formula to Wikipedias in twenty languages – the English language edition is excluded "due to its size and difficulties in processing in all dimensions", and the second and third largest Wikipedias (German and French) are missing as well. In the final "autoreferentiality index", the Icelandic, Japanese and Swahili Wikipedias come out as the most local-focused among these twenty, while, curiously, the Catalan edition which prompted the research question has the lowest autoreferentiality value.

## Does "In the news"-like attention have a positive effect on article quality?

A five page paper<sup>[5]</sup> by a Ph.D. student in Computer Science at the University of Iowa examines "The Impact of Heavy Editorial Events on Wikipedia Page Quality" – for example the flurry of edits to the article Elizabeth Taylor after the actor's death in March 2011. To measure quality, the approach of an earlier paper<sup>[6]</sup> is used, which assigns article contributors a reputation value depending on how many of their earlier contributions have been deleted, and by whom, and also takes into account whether the article revision in question was reverted later. The resulting formula was applied to "high editorial events" in 100 articles of the English Wikipedia, from the start of Wikipedia in 2001 until the beginning of 2010. As expected, the data supported the hypothesis that "high editorial events would contribute positively to a page's quality". The five articles impacted most positively among the studied sample (biased toward the beginning of the alphabet) were art, Allen Ginsberg, anarcho-capitalism, chiropractic and death. The paper also found that a higher increase in the edit rate was associated with a higher quality increase, but does not address the question of whether the relation could be explained by the mere number of edits (i.e. whether the same number of edits over a longer time might have had the same effect).



Caused a "heavy editorial event" earlier this year: Elizabeth Taylor

## Detecting POV pushing editors

A working paper posted this month to ArXiv with the title "Pushing Your Point of View: Behavioral Measures of Manipulation in Wikipedia" presents a method to score the neutrality of Wikipedia contributors and to "detect potential POV pushing behavior".<sup>[7]</sup> The authors propose two metrics to quantify an editor's involvement in controversial topics. The first metric (*Controversy score* or *C-score*) measures the amount of attention spent by an individual editor on controversial articles, where controversiality is defined on the basis of several quantitative factors previously established in the literature. The second metric (*Clustered Controversy score* or *CC-score*) quantifies the focus of an editor's attention on controversial articles on the same topic or very similar topics: the purpose of this metric is to tease apart editors involved in genuine controversy resolution (such as administrators who are likely to participate in a broad range of discussions on controversial topics) from "potentially manipulative users" who focus their attention on a narrow set of controversial topics. To assess the validity of the above metrics the authors test their discriminatory power at identifying which editors are blocked and which are regular users who were never blocked. The remainder of the paper examines the breakdown of edits by administrators immediately after a successful Request for Adminship. The results, based on qualitative coding by a single reviewer, suggests that some topical areas in the English Wikipedia (such as politics and media) are more likely to be frequently edited by administrators with a high C-score and CC-score than any other topical categories.

## Historian of encyclopedias reviews *Good Faith Collaboration*

The most recent issue of *Annals of Science* (a scholarly journal about the history of science and technology, founded in 1936) contains a four-page review<sup>[8]</sup> of Joseph Reagle's book *Good Faith Collaboration: The Culture of Wikipedia* (published in 2010 and recently released<sup>[24]</sup> on the Web under a CC-BY-NC-SA license). The reviewer Jeff Loveland<sup>[9]</sup>, who has written extensively about the early history of encyclopedias, criticizes the book for having "one major weakness, namely in historical contextualization" (he mentions two 18th-century precedents which should have been given more attention, as they, like Wikipedia, intended to include contributions from the public: Vincenzo Coronelli's *Biblioteca Universale* and Zedler's *Universal-Lexicon*) – and rejects Reagle's claim that "historically, reference works have made few claims about neutrality as a stance of collaboration, or as an end result": "References to such values as impartiality, unbiasedness and objectivity are frequent in the prefaces of encyclopaedias over the last three hundred years". On the other hand, the reviewer praises the book for "com[ing] close to offering" a comprehensive introduction to Wikipedia, "touching as it does on nearly all aspects of the encyclopaedia" and he commends the author's writing style as "informal, energetic and appropriately paced". The "insightful and worthwhile" ethnography of Wikipedia is highlighted as the second success of the book.

Regarding chapter 3<sup>[10]</sup> of the book, which postulates Neutral Point of View and Assume Good Faith as the two principles at "the heart of Wikipedia collaboration", the review recommends "Anne Goldgar's study of conduct as a force binding together the early modern Republic of Letters in *Impolite Learning* (1995) [as] an interesting point of comparison" regarding "the historical connection between knowledge and civility". Commenting on chapter 7<sup>[11]</sup>, which examines criticism of Wikipedia, Loveland observes that "the portrayal by critics of a possible Wikipedian collective intelligence as anti-individualistic, or anti-rationalistic seems opportunistic and off-the-mark. Meanwhile, Wikipedia now bears the brunt of a refurbished but centuries-old accusation against encyclopaedias, namely that they trivialize and fragment knowledge."

## Briefly

- **Automatically assessing editors' reputations:** Wöhner, Thomas and Köhler present new metrics for automatic reputation assessment of Wikipedia editors. They evaluate seven potential metrics for reputation assessment including editing frequency and contribution to high-quality articles, plus new metrics that they conceived including 'efficiency' which they define as 'the portion of an author's contribution that is persistent and quantifies the acceptance of the author within the Wikipedia community.' They evaluate the metrics using a database of the

Germany Wikipedia from January, 2008 and tested their metrics against Wikipedia's internal user classification of blocked users, administrators and anonymous users. They conclude that editing efficiency is most significant for reputation assessment since it was able to distinguish between blocked and regular authors with an accuracy rate of 86%.<sup>[12]</sup>

- **Students reflect on Wikipedia assignments:** Chen and Reber present the results of a pilot study where students from a Norwegian and a German university were asked to reflect on their experience writing a Wikipedia article as a course assignment. They provide the results of two independent judges who analyzed the written reflections of students on ten dimensions including: relevance for society, learning outcome and difficulty, among others. The authors conclude that students were highly motivated by the task and 'have learned much about the topic that they wrote about'.<sup>[13]</sup>
- **Too few newbies?:** A paper titled "Too Few New Wikipedians? Modelling Effort and Participation in Wikipedia"<sup>[14]</sup> evaluates "the efficiency of the Wikipedia projects in different languages in transforming inputs (people using the Internet) into outputs (articles). We find a decreasing return to scale in the biggest projects, but the size or the age of the projects are not the main explanation for the variations in efficiency we see."
- **How student editors use sources: synthesis vs plagiarism:** Sormunen and Lehtio report the findings of a pilot study on how Finnish secondary-school students use sources when they are required to contribute to Wikipedia as part of their coursework. They interviewed, observed, and analyzed the work of 11 groups of students, and found that: (1) the students relied almost exclusively on Web sources, (2) a sizable fraction (33%) of their work was copied verbatim (or very lightly edited) from their sources, (3) 30% of sources used were not cited at all. The dataset in this study is extremely small and the sample was not designed to be at all representative. Still, the conclusions are disconcerting, especially considering the recent controversy over student plagiarism in a related Wikipedia writing program (*Signpost* coverage: "A post-mortem on the Indian Education Program pilot"). The interviews with the students could potentially provide insight into why student plagiarism occurs.<sup>[15]</sup>
- **Tracking changes in Wikipedia:** A student thesis in Computer Science at the University of Dresden<sup>[16]</sup> describes the prototype of a software that tracks and categorizes edits on Wikipedia – trying, among other things, to detect articles that are being affected by external events. In a test sample containing articles that had been subject to major news events in recent years, such as Fukushima Daiichi Nuclear Power Plant or Dominique Strauss-Kahn, "about 74% of the events ... have been detected and about 68% of these detected events (74%) are recognized correctly."
- **Gendered language on Wikipedia:** Several Wikimedians have announced<sup>[17]</sup> a study titled "Mind the Gap(s)! Writing Styles of Female Editors on Wikipedia",<sup>[18]</sup> applying algorithms that try to classify a text as "male" or "female" (based on the frequency of "male keywords" and "female keywords") to text contributions by editors who state their gender on their user page (1,119 females and 722 males). Among the conclusions: "While the data is insufficient to reach the conclusion that Wikipedia attracts females who code their language usage as male in all circumstances on-wiki and off-wiki, we have shown that females use a more male style of writing when writing for Wikipedia."

## References

- [1] Wu, Guangyu, Martin Harrigan, and Pádraig Cunningham (2011). Characterizing Wikipedia Pages using Edit Network Motif Profiles. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents (SMUC'11) at the 20th ACM Conference on Information and Knowledge Management (CIKM'11)*, ACM Press, October 28, 2011. DOI (<http://dx.doi.org/10.1145/2065023.2065036>)
- PDF ([http://ir.ii.uam.es/smuc2011/res/papers/smuc2011\\_paper07.pdf](http://ir.ii.uam.es/smuc2011/res/papers/smuc2011_paper07.pdf)) Open access
- [2] Ribé, Marc Miquel, and Horacio Rodríguez (2011) Cultural Configuration of Wikipedia: Measuring Autoreferentiality in Different Languages. In *Proceedings of Recent Advances in Natural Language Processing*, 316-322. Hissar, Bulgaria. PDF (<http://aclweb.org/anthology-new/R/R11/R11-1044.pdf>) Open access
- [3] Ribé, Marc Miquel (2011) *Cultural configuration of Wikipedia: Measuring autoreferentiality in different languages*. Universitat Politècnica de Catalunya. PDF (<http://upcommons.upc.edu/pfc/bitstream/2099.1/11700/1/memoria.pdf>) Open access
- [4] [http://collablab.northwestern.edu/wikipedia\\_api/Wikipedia/Home.html](http://collablab.northwestern.edu/wikipedia_api/Wikipedia/Home.html)

- [5] Oliver, Corey (2011) *The Impact of Heavy Editorial Events on Wikipedia Page Quality*. **PDF** (<http://www.coreyoliver.org/research/Oliver2011.pdf>) Open access
- [6] Javanmardi and C. Lopes. Statistical Measure of Quality in Wikipedia. In: 1st Workshop on Social Media Analytics (SOMA '10), July 2010. **PDF** ([http://snap.stanford.edu/soma2010/papers/soma2010\\_18.pdf](http://snap.stanford.edu/soma2010/papers/soma2010_18.pdf)) Open access
- [7] Das, Sanmay, Allen Lavoie, and Malik Magdon-Ismael (2011). Pushing Your Point of View: Behavioral Measures of Manipulation in Wikipedia. *arXiv*, November 8, 2011. **PDF** (<http://arxiv.org/abs/1111.2092>) Open access
- [8] Loveland, Jeff (2011). Review of: Good Faith Collaboration: The Culture of Wikipedia. *Annals of Science* 68 (4) (October) 555-558. **DOI** (<http://dx.doi.org/10.1080/00033790.2011.564297>) Closed access
- [9] [http://www.artsci.uc.edu/collegemain/faculty\\_staff/profile\\_details.aspx?ePID=MjcyOTk%3D](http://www.artsci.uc.edu/collegemain/faculty_staff/profile_details.aspx?ePID=MjcyOTk%3D)
- [10] <http://reagle.org/joseph/2010/gfc/chapter-3.html>
- [11] <http://reagle.org/joseph/2010/gfc/chapter-7.html>
- [12] Wöhner, Thomas, Sebastian Köhler, and Ralf Peters (2011). Automatic Reputation Assessment in Wikipedia. In *ICIS 2011 Proceedings*. **HTML** (<http://aisel.aisnet.org/icis2011/proceedings/onlinecommunity/5>) Closed access
- [13] Chen, Weiqin, and Rolf Reber (2011). Writing Wikipedia Articles as Course Assignment. In *Proceedings of the 19th International Conference on Computers in Education*, T. Hirashima et al. (Eds). Chiang Mai, Thailand. **PDF** ([http://122.155.1.128/icce2011/program/proceedings/pdf/C6\\_S14\\_141S.pdf](http://122.155.1.128/icce2011/program/proceedings/pdf/C6_S14_141S.pdf)) Open access
- [14] Crowston, Kevin, Nicolas Jullien, and Felipe Ortega (2011) Too Few New Wikipedians? Modelling Effort and Participation in Wikipedia. SSRN eLibrary. **PDF** (<http://ssrn.com/paper=1960696>) Open access
- [15] Sormunen, Eero, and Leeni Lehtio (2011) *Authoring Wikipedia articles as an information literacy assignment – copy-pasting or expressing new understanding in one's own words?* **PDF** ([https://www12.uta.fi/blogs/know-id/files/2011/10/SormunenLehtio\\_IR2011.pdf](https://www12.uta.fi/blogs/know-id/files/2011/10/SormunenLehtio_IR2011.pdf)) Open access
- [16] Deng, Yihan (2011) *Change Tracking in Wikipedia*. Master Thesis, **PDF** ([http://www.rn.inf.tu-dresden.de/uploads/Studentische\\_Arbeiten/Belegarbeit\\_Deng\\_Yihan.pdf](http://www.rn.inf.tu-dresden.de/uploads/Studentische_Arbeiten/Belegarbeit_Deng_Yihan.pdf)) Open access
- [17] <http://www.gossamer-threads.com/lists/wiki/foundation/257912>
- [18] LauraHale, Hawkeye7, Pine and others, *Mind the Gap(s)! Writing Styles of Female Editors on Wikipedia*.

## Issue 1(6): December 2011

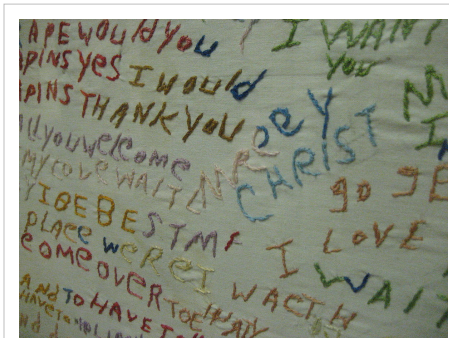
### Psychiatrists: Wikipedia better than Britannica; spell-checking Wikipedia; Wikipedians smart but fun; structured biological data

**With contributions by:** Tbayer, DarTar and Jodi.a.schneider

#### Mental health information on Wikipedia more accurate than Britannica and Kaplan & Sadock psychiatry textbook

In an article for *Psychological Medicine*,<sup>[1]</sup> ten researchers from the University of Melbourne conclude that "the quality of information on depression and schizophrenia on Wikipedia is generally as good as, or better than, that provided by centrally controlled websites, Encyclopaedia Britannica and a psychiatry textbook."

The study focused on ten mental health topics (e.g. "antidepressants and suicide in young people" or "side-effects of antipsychotics"), five each in the areas of depression and schizophrenia. "Using the topic terms (or synonyms) as key words for the searches or through manual browsing, content relating to these topics was extracted from [Wikipedia and 13 other websites selected for prominent Google results for *depression* and *schizophrenia*] and from the most recent edition of *Kaplan & Sadock's Comprehensive Textbook of Psychiatry* ... and the online version of *Encyclopaedia Britannica*" by two reviewers. For both depression and schizophrenia, three psychologists with clinical and research expertise in that area evaluated these extracts on accuracy, up-to-dateness, breadth of coverage, referencing and readability, on a scale from 1 to 5 ("e.g. Accuracy: 1 = many errors of fact or unsubstantiated opinions, 3=some



Wikipedia articles on schizophrenia and other mental health topics were assessed for accuracy, richness of references and readability.

errors of fact or unsubstantiated opinions, 5 = all information factually accurate"). As in an earlier study of the quality of health information on Wikipedia (*Signpost* coverage: "Wikipedia's cancer coverage is reliable and thorough, but not very readable"), readability was also measured using a Flesch–Kincaid readability test, which is calculated from word and sentence lengths.

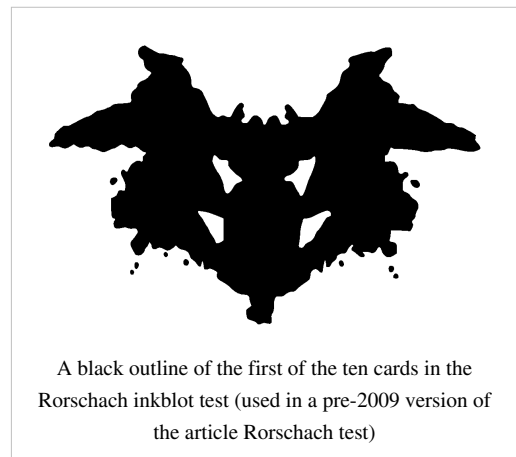
For both depression and schizophrenia, Wikipedia scored highest in the accuracy, up-to-dateness, and references categories – surpassing all other resources, including WebMD, NIMH, the Mayo Clinic and Britannica online. In breadth of coverage, it was behind Kaplan & Saddock and others for both areas. And "of the online resources, Wikipedia was rated the least readable [by the human reviewers], although some of its topics received an average rating." Likewise, the Wikipedia content had relatively high Flesch–Kincaid Grade Level indices (around 16 for schizophrenia and 15 for depression – indicating that a tertiary level of education is necessary to understand the content), similar to that of Britannica but higher than most other resources examined.

The authors note that their "findings largely parallel those of other recent studies of the quality of health information on Wikipedia" (citing eight such studies published between 2007 and 2010):

*"Despite variability in the methodologies and conclusions of these studies, the overall implication is that Wikipedia articles on health topics typically contain relatively few factual errors, although they may lack breadth of coverage. ... Given the number of patients, would-be patients and concerned others using the internet to search for information on health issues, it seems that Wikipedia is an appropriate recommendation as an information source."*

### Psychologists gauge impact of Wikipedia's Rorschach test coverage

A paper in the *Journal of Personality Assessment*<sup>[2]</sup> tried to assess the impact of the Wikipedia article Rorschach test on psychologists' use of that test. As summarized by the authors, "In the summer of 2009, an emergency room physician [User:Jmh649 - James Heilman, MD] posted images of all 10 Rorschach inkblots on [...] Wikipedia. The images were accompanied by descriptions of "common responses" to each blot. ... a fierce debate ensued between some psychologists who claimed that posting the inkblots is a threat to test security and other individuals, including some psychologists and other mental health professionals, who argued that all information should be freely available, including full details of the Rorschach". (In fact, the debates on whether to display versions of the inkblots in the article go back to at least 2005, at first accompanied by rather spurious copyright claims - Rorschach died in 1922.) The authors note that the inkblots had already been revealed to the general public in a 1980s book and cite an earlier study<sup>[3]</sup> that had found "particularly damaging information" about personality assessment tests on the Internet as early as 2000, "including examples of test stimuli from... the Rorschach" (presumably including this site<sup>[4]</sup>). Still, "Internet coverage of the Rorschach appeared to grow exponentially during" the 2009 debate about the Wikipedia article, which made it to the front page of the New York Times (*Signpost* coverage: "Rorschach test dispute reported").



The first part of the study examined the top 50 Google search results for "Rorschach"<sup>[5]</sup> (excluding "watchmen" in order to filter out results about a comic book and film) and "inkblot test"<sup>[6]</sup>, coding them into four levels representing the "threat each site presents to test security and the extent to which the content of the site might aid an individual in dissimulating on the Rorschach". 44% of the sites were classified as Level 0 ("no threat"), e.g. home page of bands with "Rorschach" in their name, and 15% as Level 1 ("minimal threat"). The 22% Level 2 ("indirect threat") sites which "tended to discuss test procedures more explicitly" apparently included "several 'official' Rorschach Web sites, where one is able to register for Continuing Education Rorschach workshops, [and which] also

allow visitors to purchase materials that contain sensitive test information. For example, certain training Web sites allow individuals to purchase training texts and instructional media without requiring a license or other professional credentials". The authors find it "disturbing" that many sites in this threat category "were authored by psychologists". 19% of the sites were classified as the highest level, "direct threat", e.g. many that contained depictions of one or more Rorschach inkblots, or specific information about how responses are interpreted. Together with results about the high percentage of Internet users consulting Wikipedia for health information (36% in the US in 2007 according to Pew research), the authors conclude that "we can no longer presume that examinees have not been exposed to this information prior to an assessment".

The second part of the study likewise starts out with a Google News search for "Rorschach" and "Wikipedia", noting that "of the 25 news stories reviewed, 13 included one or more of the Rorschach inkblots, with Card I as the most frequently displayed", and eventually arriving at five media stories about the controversy which allowed readers' comments ([7], [8], [9], [10], [11]). The altogether 520 comments on these stories were "coded according to the opinion expressed by the writer regarding each of the following categories: (a) the field of psychology, (b) psychologists, and (c) the Rorschach." While the vast majority did not state a clear opinion on the first two categories, the authors note that "Of those comments that did express an opinion toward psychologists [ca. 16%] most were overwhelmingly negative." Many more of the commenters on the Wikipedia/Rorschach news stories expressed an opinion about the test itself: "In total, 182 (35%) of comments were classified as unfavorable toward the Rorschach, whereas only 55 (11%) were coded as favorable toward the Rorschach. The remaining 283 (54%) comments were categorized as neutral or not mentioned." Among those who identified as mental health professionals, 61% expressed a favorable opinion about the test and 15% a negative one.

Asked for his comment on the paper, Heilman said: "My main criticism of their paper is that they seem to take as axiomatic that exposure to these images hurts test reliability without any real evidence to back it up. Otherwise it is an interesting piece." (The paper includes a section reviewing literature on "the impact of 'coaching' on psychological tests", however it does not mention results pertaining specifically to the Rorschach test, and mostly concerns subjects who deliberately try to "cheat" on such tests, rather than those who have accidentally been exposed to a test's material before.)

## Spell-checking the English Wikipedia

University of Nebraska-Lincoln MBA candidate Jon Stacey reports on the results of a proof-of-concept tool to measure the rate of misspelled words in the English Wikipedia over time.<sup>[13][14]</sup> A text parser (code available for download<sup>[15]</sup>) was applied to a random sample of 2,400 articles. Instead of considering the latest revision, a random revision from the history of each article was used. The final corpus was obtained by stripping markup and non-ASCII characters as well as article sections such as the references and table of contents. Words were matched against a dictionary obtained by manually combining *I2dicts* and *SCOWL* (source<sup>[16]</sup>) with Wiktionary.

The results show that the percentage of misspellings has been growing steadily, reaching 6.23% for revisions created in 2011. Several weaknesses with the method are discussed, including the lack of Unicode support, the high rate of false positives, and the possibility that the rising rate might be associated with a rise in the complexity of content. The concluding remarks speculate on how semi-automated spell-checking may support editorial work at a large scale. (Wikipedians have used lists of common misspellings for many years, also integrated in semi-automatic editing tools such as AutoWikiBrowser.)

The screenshot displays the 'LanguageTool WikiCheck' interface. At the top, there is a green header with the tool's name and logo. Below this, a search bar contains the URL 'https://en.wikipedia.org/wiki/London'. Underneath the search bar, there is a list of results for the word 'London'. The results include a definition of London, its status as a city, and its location. The tool also identifies several errors in the text, such as 'London' being misspelled as 'Londen' or 'Londun', and 'London' being used as a verb. The tool provides suggestions for corrections and links to the original text in the Wikipedia article.

In related news, the developers of an open-source multilingual proofreading application called *LanguageTool* <sup>[17]</sup> released a beta application for proofreading Wikipedia articles. *wikiCheck* <sup>[12]</sup> proofreads articles from the English and German Wikipedias based on a set of customizable syntax and grammar rules. A bookmarklet is available to access the application from a browser.

### **Wikipedians are "smart but fun", and have expertise in topics they edit**

Three researchers from Stanford University and Yahoo! Research used a novel method to construct "a data-driven portrait of Wikipedia editors", as described in a preprint currently undergoing review for publication.<sup>[18]</sup> While earlier studies relied on Wikipedians participating in surveys (and identifying themselves as such), the authors mined data from users of the Yahoo! Toolbar for Wikipedia URLs containing an `&action=submit` parameter, thereby arriving at a sample of 1900 editors of the English Wikipedia.

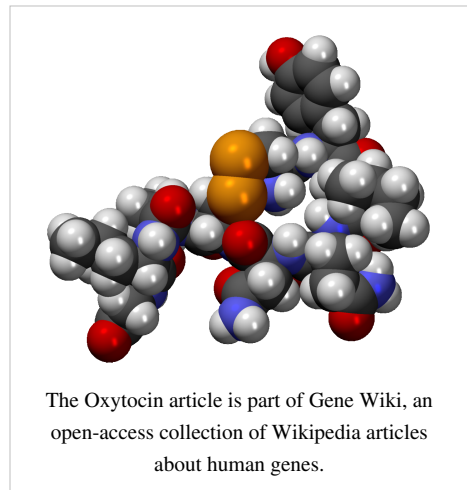
Their first main finding is that "on broad average, Wikipedia editors seem, on the one hand, more sophisticated than usual Web users, reading more news, doing more Web search, and looking up more things in dictionaries and other reference works; on the other hand, they are also deeply immersed in pop culture, spending much online time on music- and movie-related websites." However, these "entertainment lovers ... form only a highly specialized subgroup that contributes many edits".

Based on the toolbar data, the paper also tries to answer the question "Do Wikipedia editors know their domain?" and related questions, positively: "across all topical domains Wikipedia editors show significant expertise. ... We also show that more substantial edits tend to come from experts", and that logged-in editors show more expertise than IP editors. A final result is that "About half of the click chains culminating in an edit start with a Web search, with the other half originating on Wikipedia's main page."

### **Wikipedia as a database for structured biological data**

A special issue of *Nucleic Acids Research* features 11 articles describing how wikis and collaborative technology can be used to enhance biological databases. A commentary by Robert Finn, Paul Gardner and Alex Bateman<sup>[19]</sup> discusses in particular how to leverage Wikipedia, its collaborative infrastructure and large editor community to better integrate articles and biological data entries: the authors argue that the project offers an opportunity for crowdsourcing the curation and annotation of biological data, but faces major challenges for expert engagement, i.e. "how to get scientists *en masse* to edit articles" and "how to allow editors to receive credit for their work on an article".

Another article in the same issue<sup>[20]</sup> presents the Gene Wiki, an open-access and openly editable collection of Wikipedia articles about human genes. The article describes how structured data available on Gene Wiki articles is kept in sync with the data from primary databases via an automated system and how to automatically compute the quality of articles in the project at word or sentence-level using WikiTrust.



### **Individual and social drivers of participation in Wikipedia**

A thesis entitled *Individual and social motivations to contribute to Commons-based peer production* was submitted by University of Minnesota student Yoshikazu Suzuki for an MA in mass communication. The thesis presents and discusses the results from a small series of interviews as well as a survey exploring individual and social motivations of Wikipedia contributors, drawing on social identity theory, volunteerism and uses and gratifications theory. The survey, run in July 2011 with support from the Wikimedia Research Committee, collected 208 responses from a



random sample of 950 among the top English Wikipedia editors. The results, obtained by applying principal components analysis to the responses, reveal eight distinct motivational factors: providing information, the seeking of creative stimulation, concern for others' well-being, the need to be entertained, the avoidance of negative self-affect, cognitive group membership, career benefits, and social desirability. An analysis of the relative strength of each factor indicates that providing information, the seeking of creative stimulation, and concerns for others' well-being were the three strongest motivational dimensions. Grouping the eight factors into two macro-categories according to self- and other-focused motivations, the other-focused motivations were found to be significantly stronger than the self-focused motivations. The thesis reviews the implications of these results for the design of incentives for participation and editor retention. The full text of the thesis<sup>[21]</sup> and an executive summary<sup>[22]</sup> are available under open access.

### Mining article revision histories for insights into open collaboration

A paper in this month's edition of *First Monday*, ambitiously titled "Understanding collaboration in Wikipedia"<sup>[23]</sup>, reports on a statistical analysis of a complete dump of the English Wikipedia (225 million article edits) with regard to several quantities, starting with two that were introduced in a 2004 paper by Andrew Lih "as a simple measure for the reputation of [an] article within the Wikipedia": the total number of edits an article has received ("rigor") and the number of (logged-in and anonymous) users who have edited the article ("diversity"). The *First Monday* paper cites a 2007 study from the same journal, which found that featured articles tend to have more edits and contributors (while controlling for a few other variables)<sup>[24]</sup> as a justification for using "rigor" and "diversity" as proxies for article quality, but includes other quantities such as the article size change for an edit. The paper cites earlier work on evaluating Wikipedia article quality (e.g. dismissing the well-known 2005 *Nature* study based on the mistaken assumption that it had "only focused on featured articles"), but does not discuss existing attempts at more sophisticated quantitative quality heuristics.

The *First Monday* paper highlighted that if consecutive edits by the same user are counted as one, the overall number of article revisions drops by more than 33%, "revealing that one in three revisions in Wikipedia consist of users responding to their own edits or continuing an ongoing edit begun by themselves". "Article diversity" ranged up to 12,437 contributors per article, with a median of 12 and an average of 32. One of the main conclusions is that "rather than reflecting the contributions and expertise of a large group of people, the typical article in Wikipedia reflects the efforts of a relatively small group of users (median of 12) who make a relatively small number of edits (median of 21)."

Supporting the assumption that most edits do not result in significant changes in content, the study finds that 31% of all revisions cause a size change of fewer than 10 characters, and 51% a change of fewer than 30 characters, with an apparently significant peak at a 4-character difference, presumably related to the insertion or removal of the four brackets ("[[ ... ]]") that generate a wikilink.

The author notes the slight decrease in the overall number of edits since 2008, but tentatively explains it by the increasingly complete coverage of encyclopedic topics, and doesn't share the widespread concerns about declining or stagnating editor activity: "participation in Wikipedia seems to remain as healthy as ever as revisions made per article created each year has annually increased since 2001 without exception".

A different paper<sup>[25]</sup> from last year's "Collaborative Innovation Networks Conference" similarly promises far-reaching insights from "Deconstructing Wikipedia" solely based on revision history statistics without analyzing the actual content changes, using a much smaller sample – 30 featured articles from the English Wikipedia, but also

The screenshot shows the revision history for the article "Vitamin C". It displays two versions of the text side-by-side. The left version is the current revision (10451), and the right version is a previous revision (10450). Changes between the two versions are highlighted in green (additions) and red (deletions). The text discusses the chemical structure of Vitamin C, its role in the body, and its function as an antioxidant. The revision history interface includes a title bar, a list of revisions, and a comparison view.

Article revision histories are a rich data source to study patterns of collaboration on Wikipedia.

including timestamps. The data did not confirm the hypothesis that "the editor who initiated an article would have a high level of involvement in the article's creation": for only five of the 30 articles, the initial author was the most frequent contributor.

A second conclusion is that for all of the articles in the sample, "there is a single Wikipedian whose contributions far exceed all others", ranging from 8% to 82% of the articles with an average of 39% (but the analysis does not seem to have sought to quantify the extent to which this exceeds the contributions of the second most frequent contributor). The author indicates that this supports Jaron Lanier's "oracle illusion" criticism of a supposed presentation of Wikipedia as a product of "the crowd". Somewhat tautologically, the author observes "that the control of an individual editor seemed to be reduced as more editors joined the process", and points to the need to analyze "a significantly larger number of articles" to answer the question whether "too many cooks spoil the stew" (apparently unaware of the significant body of earlier literature on this subject, starting with a 2005 paper that presented an answer in its title: "Too many cooks don't spoil the broth <sup>[26]</sup>", and including the 2007 study which the above reviewed First Monday paper relied on).

A third result of the paper, which likewise might not surprise those already familiar with Wikipedia's editing processes, is that "the creation process is continuous and can go on for a very long time", with even articles about historic events from the distant past continuing to receive edits.

The author, an assistant professor in management and marketing at Virginia State University, concludes the paper by urging his readers to start "thinking about how the wiki platform, itself, is influencing the creation process".

## Briefly

- The Wikimedia Research Committee launched a public consultation <sup>[27]</sup> on the future **data/research infrastructure for Wikimedia**, in an effort to understand how to best serve the research and developer community with open data from our projects. The consultation will remain open through January 2012 and the full set of responses will be shared under a CC0 license.
- **Semantic enhancements:** In "Enhancing Wikipedia with semantic technologies"<sup>[28]</sup>, Lee et al. review existing interfaces for semantic search and present their own platform for enhancements. Based on small-scale user tests, they find that one of their three enhancements – range-based queries – are strongly preferred by users, who would find them desirable not only in Wikipedia but on the wider web. A longer summary is available on AcaWiki <sup>[29]</sup>.
- **English and Finnish Wikipedias egalitarian, Japanese hierarchical:** A paper titled "Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language Wikipedias"<sup>[30]</sup> (from the *2010 Collaborative Innovation Networks Conference*, as was the revision statistics paper reviewed above) applied social network analysis to collaboration on featured articles on the English, German, Japanese, Korean, and Finnish Wikipedias. It "found notable differences in the communication behavior among egalitarian cultures such as the Finnish, and quite hierarchical ones such as the Japanese. While the English language Wikipedia shows a distinctive pattern, most likely because it is by far the largest and frequently exploring new concepts copied by others, it seems to follow more the Finnish egalitarian, than the Japanese hierarchical style".
- User:Emijrp shared a link on wiki-research-l <sup>[31]</sup> listing **2,596 scholarly references on wikis** <sup>[32]</sup>, obtained by scraping Google Scholar results (on December 22, 2011), as part of a project to build a comprehensive bibliography about wikis – a challenging task that has seen various earlier attempts and was the subject of a workshop at this year's WikiSym (see the October and April editions of this research report).
- **Should doctors use and edit Wikipedia?:** An editorial in the *Journal of the Royal Society of Medicine* <sup>[33]</sup> asks whether doctors should reject the use of Wikipedia. The two page article (one-day access: US\$30.00) cites some results about the popularity of Wikipedia among medical students, young physicians and the general public, and for some reason highlights the malicious edits of British journalist Johann Hari as example for the downsides of Wikipedia's free editability. It contains a review of the literature on the reliability of Wikipedia's medical

information which is less thorough than that of the *Psychological Medicine* article reviewed above, and comes to a less approving but still somewhat positive conclusion: "Although Wikipedia entries are often poorly structured and difficult to understand, they are comparable in accuracy to some online resources, such as health insurance websites." In the end, the authors seems to lean towards recommending against ignoring Wikipedia: "One risk of clinicians disengaging from Wikipedia is that only contributors motivated by personal experience (e.g. patient anecdote) or vested interests (e.g. individual clinicians, institutions or companies promoting their own ideas and products) will remain."

## References

- [1] Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., Nelson, B., et al. (2011). Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine*, pp. 1-10. **DOI** (<http://dx.doi.org/10.1017/S003329171100287X>) Closed access
- [2] Schultz, D. S., & Loving, J. L. (2012). Challenges Since Wikipedia: The Availability of Rorschach Information Online and Internet Users' Reactions to Online Media Coverage of the Rorschach–Wikipedia Debate. *Journal of Personality Assessment*, 94(1), 73-81. Routledge. **DOI** (<http://dx.doi.org/10.1080/00223891.2011.627963>) Closed access
- [3] Ruiz, M., Drake, E., Glass, A., Marcotte, D., & van Gorp, W. (2002). Trying to beat the system: Misuse of the Internet to assist in avoiding the detection of psychological symptom dissimulation. *Professional Psychology: Research and Practice*, 33, 294–299 **PDF** (<http://www.sciencedirect.com/science/article/pii/S0735702802000400>) Closed access
- [4] <http://web.archive.org/web/20001001065937/http://deltabravo.net/custody/roorschach.htm>
- [5] <http://www.google.com/search?q=Rorschach+-+watchmen>
- [6] <http://www.google.com/search?q=%22inkblot%20test%22>
- [7] <http://www.digitaljournal.com/article/276688>
- [8] <http://www.findingdulcinea.com/news/science/2009/july/Rorschach-Fail-The-Test-s-Validity-Is-Again-Scrutinized-as-Plates-Appear-on-Wikipedia.html>
- [9] <http://io9.com/5344390/all-of-roorschachs-secrets-revealed>
- [10] <http://www.cbc.ca/technology/story/2009/07/31/roorschach-test.html>
- [11] <http://www.nytimes.com/2009/07/29/technology/internet/29inkblot.html?r=1>
- [12] <http://community.languagetool.org/wikiCheck>
- [13] Stacey, Jon (2011). *Text mining Wikipedia for misspelled words*. **HTML** (<http://jonsview.com/text-mining-wikipedia-for-misspelled-words>) Open access
- [14] Fogarty, Kevin (December 23, 2011). Wikipedia test showes Americans' ability too spel is detereorating (<http://www.itworld.com/internet/235523/wikipedia-test-showes-americans-ability-too-spel-detereorating>). ITworld.
- [15] [http://jonsview.com/text-mining-wikipedia-for-misspelled-words#appendices\\_and\\_references](http://jonsview.com/text-mining-wikipedia-for-misspelled-words#appendices_and_references)
- [16] <http://wordlist.sourceforge.net/>
- [17] <http://www.languagetool.org/>
- [18] West, Robert, Ingmar Weber, and Carlos Castillo (2011). *Smart but Fun: A Data-Driven Portrait of Wikipedia Editors*. **PDF** (<http://ai.stanford.edu/~west1/pubs/wikiedits.pdf>) Open access
- [19] Finn, Robert D, Paul P Gardner, and Alex Bateman (2011). Making your database available through Wikipedia: the pros and cons. *Nucleic acids research*, 40(1) **DOI** (<http://dx.doi.org/10.1093/nar/gkr1195>) • **HTML** (<http://nar.oxfordjournals.org/content/40/D1/D9.full>) Open access
- [20] Good, Benjamin M, Erik L Clarke, Luca de Alfaro, and Andrew I Su (2011). The Gene Wiki in 2011: Community intelligence applied to human gene annotation. *Nucleic acids research* 40 (1): D1255-1261. **DOI** (<http://dx.doi.org/10.1093/nar/gkr925>) • **HTML** (<http://nar.oxfordjournals.org/content/40/D1/D1255.full>) Open access
- [21] Suzuki, Yoshikazu (2011) *Individual and social motivations to contribute to Commons-based peer production*, MA thesis, University of Minnesota. **PDF** ([http://conservancy.umn.edu/bitstream/119040/1/Suzuki\\_Yoshikazu\\_November2011.pdf](http://conservancy.umn.edu/bitstream/119040/1/Suzuki_Yoshikazu_November2011.pdf)) Open access
- [22] <http://www.scribd.com/doc/74939326/>
- [23] Kimmons, Royce (2011). "Understanding collaboration in Wikipedia". *First Monday* 16 (12). **HTML** (<http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3613/3117>) Open access
- [24] Wilkinson, D.M., and B.A. Huberman (2007). "Assessing the value of cooperation in Wikipedia". *First Monday* 12 (4). **PDF** (<http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1763/1643>) Open access
- [25] Feldstein, A. (2011). "Deconstructing Wikipedia: Collaborative Content Creation in an Open Process Platform". *Procedia – Social and Behavioral Sciences*, 26, 76–84. **DOI** (<http://dx.doi.org/10.1016/j.sbspro.2011.10.564>) Open access
- [26] <http://www.mediachange.ch/publications/27/>
- [27] [http://docs.google.com/spreadsheets/viewform?hl=en\\_US&formkey=dGNBSGFUcTdJLUxLcGpWUNoQXM0SGc6MQ](http://docs.google.com/spreadsheets/viewform?hl=en_US&formkey=dGNBSGFUcTdJLUxLcGpWUNoQXM0SGc6MQ)
- [28] Lian Hoy Lee, Christof Lutteroth, and Gerald Weber (2011). Enhancing Wikipedia with Semantic Technologies. In *iUBICOM'11: The 6th International Workshop on Ubiquitous and Collaborative Computing*, 2011. **PDF** ([http://www.bcs.org/upload/pdf/ewic\\_ubi11\\_paper3](http://www.bcs.org/upload/pdf/ewic_ubi11_paper3)).

- pdf) Open access
- [29] [http://acawiki.org/index.php?title=Enhancing\\_Wikipedia\\_with\\_semantic\\_technologies](http://acawiki.org/index.php?title=Enhancing_Wikipedia_with_semantic_technologies)
- [30] Nemoto, Keiichi, and Peter A. Gloor (2011). Analyzing Cultural Differences in Collaborative Innovation Networks by Analyzing Editing Behavior in Different-Language Wikipedias. *Procedia - Social and Behavioral Sciences* 26 (January 2011): 180-190. **DOI** (<http://dx.doi.org/10.1016/j.sbspro.2011.10.574>) Open access
- [31] <http://lists.wikimedia.org/pipermail/wiki-research-l/2011-December/001753.html>
- [32] <http://pastebin.com/6MzvR6Vi>
- [33] Metcalfe, D., & Powell, J. (2011). Should doctors spurn Wikipedia? *JRSM*, 104 (12), 488-489. **PDF** (<http://dx.doi.org/10.1258/jrsm.2011.110227>) Closed access
-

# Article Sources and Contributors

**About** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3204576> *Contributors:* DarTar, Peteforsyth, Rock drum, Tbayer (WMF), Trijnstel, 3 anonymous edits

**Issue 1(1): July 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3041953> *Contributors:* Angr, DarTar, Ebe123, HaeB, Jarry1250, Jtmorgan, Mietchen, Nabla, Peteforsyth, Skomorokh, Steven (WMF), Tbayer (WMF), 1 anonymous edits

**Issue 1(2): August 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3041958> *Contributors:* DarTar, Graham87, HaeB, John Broughton, Mietchen, Peteforsyth, Seb az86556, Tbayer (WMF), Tony1, Wavelength

**Issue 1(3): September 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3041975> *Contributors:* Cuprum17, DarTar, EpochFail, Graham87, HaeB, JoJan, Jodi.a.schneider, Jon Harald Soby, Mietchen, Peteforsyth, Qwfp, SMasters, Tbayer (WMF), Tony1

**Issue 1(4): October 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3041956> *Contributors:* Boghog, DarTar, ElKeVbo, Hewhoamareismyself, Jarry1250, Jodi.a.schneider, Marcus Qwertyus, Peteforsyth, Phoebe, Ruud Koot, SMasters, Skomorokh, Tbayer (WMF), Tpbradbury

**Issue 1(5): November 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3119811> *Contributors:* DarTar, Graham87, Hfordsa, Jack Greenmaven, Rcsprinter123, Romanesco, Skomorokh, Staciou, Tbayer (WMF), Utar, Wavelength, 3 anonymous edits

**Issue 1(6): December 2011** *Source:* <http://meta.wikimedia.org/w/index.php?oldid=3190149> *Contributors:* DarTar, Jodi.a.schneider, Kaldari, Lambiam, SMasters, Tbayer (WMF), Tony1

# Image Sources, Licenses and Contributors

**File:Wikimedia Research Newsletter.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Wikimedia\\_Research\\_Newsletter.jpg](http://meta.wikimedia.org/w/index.php?title=File:Wikimedia_Research_Newsletter.jpg) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* DarTar, Rock drum

**File:Feed-icon.svg** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Feed-icon.svg> *License:* unknown *Contributors:* unnamed (Mozilla Foundation)

**File:Open Access logo PLoS transparent.svg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Open\\_Access\\_logo\\_PLoS\\_transparent.svg](http://meta.wikimedia.org/w/index.php?title=File:Open_Access_logo_PLoS_transparent.svg) *License:* Creative Commons Zero *Contributors:* art designer at PLoS, modified by Wikipedia users Nina, Beao, and JakobVoss

**File:Closed Access logo alternative.svg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Closed\\_Access\\_logo\\_alternative.svg](http://meta.wikimedia.org/w/index.php?title=File:Closed_Access_logo_alternative.svg) *License:* unknown *Contributors:* Jakob Voß, influenced by original art designed at PLoS, modified by Wikipedia users Nina and Beao

**File:Obama discussion tree.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Obama\\_discussion\\_tree.jpg](http://meta.wikimedia.org/w/index.php?title=File:Obama_discussion_tree.jpg) *License:* Creative Commons Attribution-Sharealike 2.0 *Contributors:* David Laniado, Riccardo Tasso, Yana Volkovich, Andreas Kaltenbrunner

**File:Deletion process on English Wikipedia (flowchart).jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Deletion\\_process\\_on\\_English\\_Wikipedia\\_\(flowchart\).jpg](http://meta.wikimedia.org/w/index.php?title=File:Deletion_process_on_English_Wikipedia_(flowchart).jpg) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Jodi.a.schneider

**File:HelpRequestLocations.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:HelpRequestLocations.png> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Jtmorgan

**File:HelpRequestTypes.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:HelpRequestTypes.png> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Jtmorgan

**File:Edits to English Wikipedia trending articles.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Edits\\_to\\_English\\_Wikipedia\\_trending\\_articles.png](http://meta.wikimedia.org/w/index.php?title=File:Edits_to_English_Wikipedia_trending_articles.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Drdee

**File:Trending article edits, excludes Semi-protected pages.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Trending\\_article\\_edits\\_excludes\\_Semi-protected\\_pages.png](http://meta.wikimedia.org/w/index.php?title=File:Trending_article_edits_excludes_Semi-protected_pages.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Drdee

**File:Patrol months.per user.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Patrol\\_months.per\\_user.png](http://meta.wikimedia.org/w/index.php?title=File:Patrol_months.per_user.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Steven (WMF), WereSpielChequers

**File:Patrol months.top 50.per user.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Patrol\\_months.top\\_50.per\\_user.png](http://meta.wikimedia.org/w/index.php?title=File:Patrol_months.top_50.per_user.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Steven (WMF), WereSpielChequers

**File:RevDelete Special-RevisionDelete (narrow).png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:RevDelete\\_Special-RevisionDelete\\_\(narrow\).png](http://meta.wikimedia.org/w/index.php?title=File:RevDelete_Special-RevisionDelete_(narrow).png) *License:* Public Domain *Contributors:* FT2

**File:BotanicalGardensGPOPerth.jpg** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:BotanicalGardensGPOPerth.jpg> *License:* Public Domain *Contributors:* Codeispoetry, Gobeirne, Matthiasb, Mietchen

**File:H1N1 influenza virus.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:H1N1\\_influenza\\_virus.jpg](http://meta.wikimedia.org/w/index.php?title=File:H1N1_influenza_virus.jpg) *License:* Public Domain *Contributors:* Original uploader was PigFlu Oink at en.wikipedia

**File:Wikipedia-logo-zh.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Wikipedia-logo-zh.png> *License:* logo *Contributors:* Nohat and Shizhao

**File:Revert\_effect\_on\_boldness.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Revert\\_effect\\_on\\_boldness.png](http://meta.wikimedia.org/w/index.php?title=File:Revert_effect_on_boldness.png) *License:* Creative Commons Attribution 3.0 *Contributors:* Aaron Halfaker

**File:WikiTrip egyptian revolution screenshot.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:WikiTrip\\_egyptian\\_revolution\\_screenshot.png](http://meta.wikimedia.org/w/index.php?title=File:WikiTrip_egyptian_revolution_screenshot.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* DarTar

**File:PSBOCT.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:PSBOCT.png> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Jmh649

**File:WikiSym 2011 panel.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:WikiSym\\_2011\\_panel.jpg](http://meta.wikimedia.org/w/index.php?title=File:WikiSym_2011_panel.jpg) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Tbayer (WMF)

**File:Lovastatin.svg** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Lovastatin.svg> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Panoramix303

**File:Wiki Participation Challenge.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Wiki\\_Participation\\_Challenge.png](http://meta.wikimedia.org/w/index.php?title=File:Wiki_Participation_Challenge.png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:DarTar

**File:Wikipedia-logo-v2-pl.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Wikipedia-logo-v2-pl.png> *License:* logo *Contributors:* Wikimedia Foundation

**File:Claw.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Claw.png> *License:* GNU Free Documentation License *Contributors:* User:Arbor

**File:Elizabeth Taylor trailer.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Elizabeth\\_Taylor\\_trailer.jpg](http://meta.wikimedia.org/w/index.php?title=File:Elizabeth_Taylor_trailer.jpg) *License:* Public Domain *Contributors:* MachoCarioca

**File:Cloth embroidered by a schizophrenia sufferer.jpg** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:Cloth\\_embroidered\\_by\\_a\\_schizophrenia\\_sufferer.jpg](http://meta.wikimedia.org/w/index.php?title=File:Cloth_embroidered_by_a_schizophrenia_sufferer.jpg) *License:* Creative Commons Attribution 2.0 *Contributors:* cometstarmoon

**Image:inkblot.svg** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:Inkblot.svg> *License:* Public Domain *Contributors:* Bryan Derksen, Dorgan, Drugonot, Liftarn, LjL, Rbenedict, Sarefo, 1 anonymous edits

**File:WikiCheck.jpg** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:WikiCheck.jpg> *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* DarTar

**File:OxitocinaCPK3D.png** *Source:* <http://meta.wikimedia.org/w/index.php?title=File:OxitocinaCPK3D.png> *License:* Creative Commons Zero *Contributors:* MindZiper

**File:History comparison example.png** *Source:* [http://meta.wikimedia.org/w/index.php?title=File:History\\_comparison\\_example.png](http://meta.wikimedia.org/w/index.php?title=File:History_comparison_example.png) *License:* unknown *Contributors:* User:J.smith

# License

---

Creative Commons Attribution-Share Alike 3.0 Unported  
[//creativecommons.org/licenses/by-sa/3.0/](https://creativecommons.org/licenses/by-sa/3.0/)

---