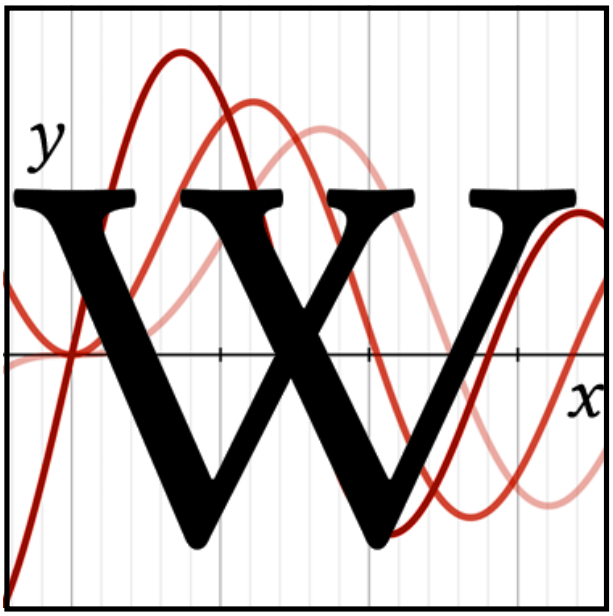
Wikimedia Research Newsletter

Volume 2 (2012)



Contents

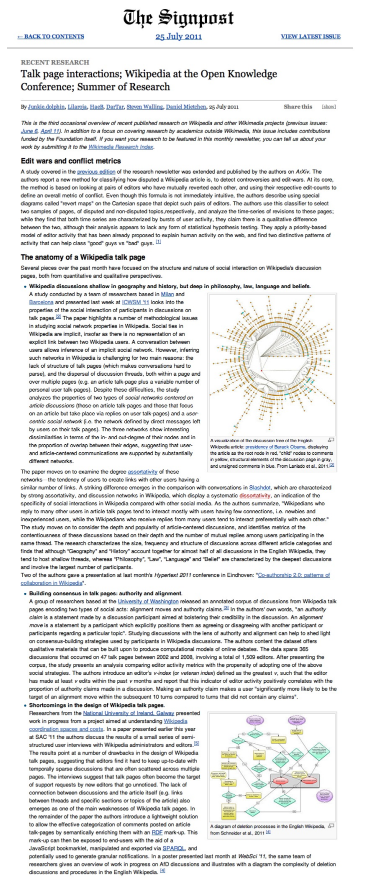
Articles

References

Article Licenses

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**.


The Signpost

← BACK TO CONTENTS | 25 July 2011 | VIEW LATEST ISSUE

RECENT RESEARCH

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

By Junkie.dolphin, Lilaroja, HaeB, DarTar, Steven Walling, Daniel Mietchen, 25 July 2011 Share this [show]

This is the third occasional overview of recent published research on Wikipedia and other Wikimedia projects (previous issues: June 6, April 11). In addition to a focus on covering research by academics outside Wikimedia, this issue includes 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.

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, 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 describe using special diagrams called "revert maps" on the Cartesian space that depict such pairs of editors. The authors use this classifier to select two samples of pages, of disputed and non-disputed topics,respectively, 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 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. [1]

The anatomy of a Wikipedia talk page

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

- **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 looks into the properties of the social interaction of participants in discussions on talk pages.[2] 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 inference of an implicit social network. However, inferring such networks in Wikipedia is challenging for two main reasons: the lack of structure of talk pages (which makes conversations hard to parse), and the dispersal of discussion threads, both within a page and over multiple pages (e.g. an article talk-page plus a variable number of personal user talk-pages). Despite these difficulties, the study analyzes the properties of two types of *social networks centered on article discussions* (those 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 show interesting dissimilarities in terms of the in- and out-degree of their nodes and in the proportion of overlap between their edges, suggesting that user- and article-centered communications are supported by substantially different networks.

A visualization of the discussion tree of the English Wikipedia article: presidency of Barack Obama, displaying the article as the root node in red, "child" nodes to comments in yellow, structural elements of the discussion page in gray, and unsigned comments in blue. From Laniado et al., 2011.[2]

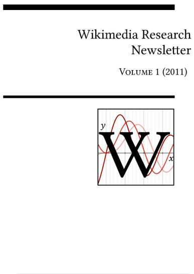
The paper moves on to examine the degree assortativity of these networks—the tendency of users to create links with other users having a similar number of links. A striking difference emerges in the comparison with conversations in Slashdot, which are characterized by strong assortativity, and discussion networks in Wikipedia, which display a systematic dissortativity, an indication of the specificity of social interactions in Wikipedia compared with 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. The research 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 number of participants.
Two of the authors gave a presentation at last month's *Hypertext 2011* conference in Eindhoven: "Co-authorship 2.0: 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 encoding two types of social acts: alignment moves and authority claims.[3] 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 to shed light on consensus-building strategies used by participants in Wikipedia discussions. The authors content the dataset offers qualitative materials that can be built upon to produce computational models of online debates. The data spans 365 discussions that occurred on 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. 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 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.[5] 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 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 the effective categorization of comments posted on article talk-pages by semantically enriching them with an RDF mark-up. This mark-up 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. [4]

A diagram of deletion processes in the English Wikipedia, from Schneider et al., 2011 [4]

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

Facts and figures



The inaugural issue of the WRN was published on July 25, 2011 – shortly after the announcement of the Wikimedia Research Index and after two Signpost articles covering recent Wikimedia research.

The six issues published in the **first volume** (July-December 2011), featuring 87 unique publications, are available as a downloadable 45-page PDF [2], and a print version can be ordered from Pediapress. The full list of publications reviewed or covered in the Newsletter in 2011 can be browsed online [3] or downloaded (as a BibTeX [4], RIS [5], PDF [6] file or in other formats), ready to be imported into reference managers or other bodies of wiki research literature.Read more... [7]

How to subscribe

- To receive the full text of each new issue in the form of an **HTML email**, sign up here [8].
- You can also subscribe to the newsletter on the Wikimedia Foundation's blog [9] via the following RSS feed:
- The table of contents of each issue is cross-posted to the **wiki-research-l** [10] mailing list.
- Follow the **@WikiResearch** feed on Identi.ca [11] or Twitter [12]. In addition to the monthly announcement of each new WRN issue, it also points to new preprints, papers or research-related blog posts before they are reviewed more fully in the upcoming issue.
- The Newsletters are also included in the weekly Wikipedia Signpost newspaper, so if you subscribe to the Signpost, you'll receive the newsletter with your regular Signpost delivery to your Wikipedia talk page.

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 [13], 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-l [10] (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.

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** [14] .

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** [15] .

Archives

Volume 3 (2013)

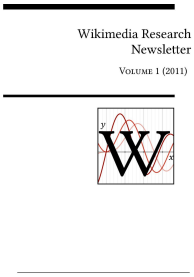
- WRN 3(1) – January 2013

Volume 2 (2012)

- WRN 2(12) – December 2012
- WRN 2(11) – November 2012
- WRN 2(10) – October 2012
- WRN 2(9) – September 2012
- WRN 2(8) – August 2012
- WRN 2(7) – July 2012
- WRN 2(6) – June 2012
- WRN 2(5) – May 2012
- WRN 2(4) – April 2012

- WRN 2(3) – March 2012
- WRN 2(2) – February 2012
- WRN 2(1) – January 2012

Volume 1 (2011)



- 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 [16]

References

[1] irc://irc.freenode.net/wikimedia
[2] https://upload.wikimedia.org/wikipedia/commons/a/a7/WRN_2011.pdf
[3] http://www.citeulike.org/user/WRN/tag/wrn2011
[4] http://www.citeulike.org/bibtex/user/WRN/tag/wrn2011
[5] http://www.citeulike.org/ris/user/WRN/tag/wrn2011
[6] http://www.citeulike.org/pdf_export/user/WRN/tag/wrn2011?citation_format=plain&file_format=pdf&q=
[7] http://blog.wikimedia.org/2012/03/16/wikimedia-research-newsletter-first-volume-new-features/
[8] https://lists.wikimedia.org/mailman/listinfo/research-newsletter
[9] https://blog.wikimedia.org/c/research-2/wikimedia-research-newsletter/
[10] http://lists.wikimedia.org/mailman/listinfo/wiki-research-l
[11] https://identi.ca/wikiresearch
[12] https://twitter.com/#!/wikiresearch
[13] http://etherpad.wmflabs.org/pad/p/WRN201302
[14] http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2764/3301
[15] http://dx.doi.org/10.1145/1998076.1998190
[16] mailto:researchnews@wikimedia.org

Issue 2(1): January 2012

Language analyses examine power structure and political slant; Wikipedia compared to commercial databases

With contributions by: Tbayer and Piotrus

Admins influence the language of non-admins

An Arxiv preprint titled "Echoes of power: Language effects and power differences in social interaction"[1] looks at the language used by Wikipedia editors. The authors look at how conversational language can be used to understand power relationships. The research analyzes how much one adapts their language to the language of others involved in a discussion (the process of language coordination). The findings indicate that the more such adoption occurs, the more deferential one is. The authors find that editors on Wikipedia tend to coordinate (language-wise) more with the administrators than with non-administrators. Further, the study suggests that one's ability to coordinate language has an impact on one's chances to become an administrator: the admin-candidates who do more language coordination have a higher chance of becoming an administrator than those who don't change their language. Once a person is elected an administrator, they tend to coordinate less.

A blog post on the website of *Technology Review* summarized the results using the headline "Algorithm Measures Human Pecking Order [2]," and highlighted the fact that one of the authors is Jon Kleinberg, known as inventor of the HITS algorithm (also known as "hubs and authorities").

Can Wikipedia replace commercial biography databases?

An article[3] by a librarian and professor at California State University offers a comparison of "biographical content for literary authors writing in English" between Wikipedia, "the web" (i.e. top Google search results) and two commercial databases: the Biography Reference Bank (BRB, now part of EBSCO Industries) and Contemporary Authors Online, motivated by the decision of the author's institution to cancel its subscription to the latter database (CAO) during a budget crisis in 2008-2009, which among other reasons had been accompanied by "a comment that this information is 'on the web'".

California State University, East Bay: Could it rely on biographical information from Wikipedia and the web alone?

The paper starts out with a literature review on the reliability of Wikipedia and then describes how the author compiled a list of 500 authors [4] (mostly from the US and UK) by "examining curricula and textbooks from English literature courses across the USA" and soliciting [5] additional suggestions from peers. These names were then searched on BRB, CAO (as part of the Literature Resource Center), Wikipedia and Google.

Regarding breadth of coverage, only six of the 500 names were "absent" on Wikipedia (meaning that they had "no entry of their own or reference in any other entry"), compared to 14 for LRC, and 50 for the Biography Reference Bank.

While the study does not seem to have attempted a systematic comparison of factual accuracy, it observes that Wikipedia "entries are less uniform than those in commercial databases. The biographical information ranges from extensive to perfunctory".

The author remarks favorably on Wikipedia's searchability:

> *"The databases and Wikipedia deal better than the Web with variant names, pseudonyms, and names that apply to multiple people. Cross-referencing is very good. [...] Wikipedia searching is very easy.There were even cases where it was easier to search Wikipedia than the databases. [...] Wikipedia also 'disambiguates' names and offers quick descriptions to enable the searcher to find the correct individual."*

A large part of the comparison consists of examining each resource's production process. Wikipedians may find parallels to their policies on biographies of living people, self-published sources and notability in the description for the Biography Reference Bank:

> *"Current Biography [the main content source of BRB] articles rely on secondary sources, but Wilson [the then publisher] has occasionally spoken directly with subjects or their proxies. Upon publication, many articles have been sent to subjects for review before being updated for the print annual and the databases. If subjects raise objections, misinformation is corrected, but not matters of public record. Adjustments may be made for privacy, for example omitting the specific names of children.*
>
> *"To be included in World Authors [another source of BRB], authors must have published more than one critically acclaimed book. [...]"*
>
> *"For autobiographies, Wilson attempted to contact subjects in Junior Authors and World Authors for a statement, but not subjects in Current Biography. [... An example offered by a Wilson employee:] For some reason, Jennie Tourel, a Russian-American opera singer, often provided false information, but, according to the Wilson biography, "passports and other documents that surfaced soon after her death helped to correct some of these inaccuracies'".*

In the conclusion, the author answers the initial question by recommending that her employer "re-subscribe to a commercial biographical database" if the budget would permit it again, because "Commercial databases provide a foundation with authoritative core content authenticated prior to publication and integrated with the fabric of information in the library's holdings. They are easy to search and reliable, although they cannot be as current as Wikipedia or the Web because of their authentication processes. Wikipedia become[sic] more impressive as searching proceeded. The focus may be on verifiability rather than authority and there may be challenges in securing contributors, but the current contributors provide citations and often include unique information." All in all she seems to favor Wikipedia and the two databases over "The web" (Google results) which "may have plenty of dross and be less reliable, harder to search, and focused on commercialism, but there are gold nuggets." She worries: "What will happen if contributors to Wikipedia and the web have no authoritative databases to use as sources?"

Students predict connections between Wikipedians

Among the student projects in a class on "Computational Analysis of Social Processes [6]" at Rensselaer Polytechnic Institute, three analyzed social networks of Wikipedia editors:

- The write-up for a project titled "Interaction vs. Homophily in Wikipedia Administrator Selection"[7] provides an analysis of factors related to one's participation (or lack thereof) in the Request for Adminship discussions. It confirms previous findings that many participants are drawn to the discussions by their personal contacts and experiences with others. The paper tries to analyze the impact of direct past interaction versus homophily (roughly defined as shared interests). The findings suggest that homophily plays a much smaller role compared to past interactions. Overall, it appears that administrators are often elected (or opposed) not by the community at large, but by a group of their closest peers. To quote from the conclusion of the paper: "This raises questions about the robustness of Wikipedia's administrator selection process which is then comprised of a very small interaction-selected group of editors."
- Another project write-up titled "Link Prediction Analysis in the Wikipedia Collaboration Graph"[8] tested various models to predict the strength of the connection between two Wikipedia editors in a "dynamic collaboration graph" that measures, at a given point in time, how often they recently edited the same page, with more recent edits weighing stronger.

- A third student paper titled "Link prediction on a Wikipedia dataset based on triadic closure"[9] likewise tested various models on a similar graph consisting of Wikipedia users as vertices, regarding the closure of triangles (i.e. if user A is connected with B, and B with C, is A connected with C as well?). Among the conclusions is that such "triadic closure, while still occurring in Wikipedia, is happening at a slower pace now than before–likely due to the influx of less active editors".

Language analysis finds Wikipedia's political bias moving from left to right

A study presented [10] earlier this month at the annual meeting of the American Economic Association which is to appear in *The American Economic Review*[11] sets out to test whether the English Wikipedia is truly neutral, by measuring bias within a sample of 28,000 entries about US political topics, examined over a decade. The bias is identified through detecting the use of language specific to one side of the American political scene (Democrats or Republicans). To quote from the article: "In brief, we ask whether a given Wikipedia article uses phrases favored more by Republican members or by Democratic members of Congress" (in the text of the 2005 Congressional Record, using a method developed in an earlier paper by Gentzkow and Shapiro who applied it to newspapers). The authors identified, as of January 2011, 70,668 articles related to US politics, about 40% of which had a statistically significant bias. They find that Wikipedia articles are often biased upon creation, and that this bias rarely changes. Early on in Wikipedia's history, most had a pro-Democratic bias, and while "by the last date, Wikipedia's articles appear to be centered close to a middle point on average", this is simply an effect of a larger amount of new pro-Republican articles than due to the existing ones having been rewritten neutrally.

While the authors made efforts to exclude articles not pertinent to US politics (requiring the terms "United States" or "America" to appear at least three times in the article text), the sample also includes the clearly international article Iraq War. And in what Wikipedians may call out as systemic bias, the authors never question their assumption that for an international encyclopedia, a lack of bias would be indicated by the replication of the spectrum of opinions present in the US Congress. As early as 2006 [12], Jimmy Wales objected to such notions with respect to the community of contributors: "If averages mattered, and due to the nature of the wiki software (no voting) they almost certainly don't, I would say that the Wikipedia community is slightly more liberal than the U.S. population on average, because we are global and the international community of English speakers is slightly more liberal than the U.S. population. ... The idea that neutrality can only be achieved if we have some exact demographic matchup to [the] United States of America is preposterous." Nevertheless, even if one turns the study on its head and reads it as a statement on average American political opinion compared to the rest of the world as reflected in the English Wikipedia, its results remain remarkable.

Briefly

- **Calls for papers** have appeared this month for
  - WikiSym 2012 [13], the eighth instance of this annual research conference on wikis and open collaboration
  - Wikimania 2012 [14], the eighth annual global conference of Wikimedians
  - Wikipedia Academy: Research and Free Knowledge [15], a conference organized by Wikimedia Germany
  - "Academic research into Wikipedia: Beyond English Wikipedia and towards comparative perspectives [16]", an upcoming issue of the e-journal *Digithum*
- **New effort at comprehensive wiki research literature database**: Wikipedian emijrp has announced [17] the launch of WikiPapers [18], a Semantic MediaWiki-based wiki dedicated to the "compilation of resources (conference papers, journal articles, theses, books, datasets and tools) focused on the research of wikis". The task of creating such a database has seen several efforts before and its difficulties were explored in a well-attended workshop at last year's WikiSym conference (see the October issue of this newsletter). Researcher Finn Årup Nielsen (who last year published an overview of such literature,[19] mentioning well over 1000 publications) pointed out [20] the possibility of exchanging content between the new wiki, the existing (likewise Semantic

MediaWiki-based) Acawiki and his own Brede Wiki.

- **Review of *Good Faith Collaboration***: Sociological journal *The Information Society* reviewed[21] Joseph Reagle's 2010 book *Good Faith Collaboration: The Culture of Wikipedia* (which was recently released [22] online under a Creative Commons license), praising it as "an accurate account of this sociocultural and sociotechnological phenomenon that Wikipedia is". The reviewer calls Wikipedia a "virtual tool and reference jim-dandy [which] is another flashpoint in our path of social anxieties" and holds that "nobody can think of a true rival [of Wikipedia] in this knowledge contest. The sins are there: lightness, temporary reliability, questionable scholarly approaches, sometimes oversimplification, sometimes data excess; however, these are venial sins and easily absolved." Somewhat cryptically, he observes that "the European Union is trying to adapt this part of Western academia to the global university system (though inevitably Anglo-American inspired)". He commends the book as an "accessible analysis [which] makes it clear that Wikipedia is not wasted knowledge; it is human thirst for knowledge and we are simply gathering scattered pieces". Gentle criticism includes that "though [...] inaccuracies are stated, two other important worries — that it is not financially sustainable, and that Wikipedia has lost touch with its founding ideal—are not as openly dealt with".
- **Predicting categories from links**: In a paper titled "Using Network Structure to Learn Category Classification in Wikipedia"[23] (the write-up of a class project for an Autumn 2011 Stanford course titled "Social and Information Network Analysis" [24]), three students describe the construction of a classifier algorithm that tries to predict from an article's ingoing and outgoing wikilinks whether it is a member of the Category:American actors – "We chose this particular category because it is one of the largest on Wikipedia (almost 25,000 pages)".
- **Wikipedia vs. library catalogue**: An article in *Library and Information Research* titled "Searching where for what: A comparison of use of the library catalogue, Google and Wikipedia"[25] analyzed search queries from users of Google (using Hitwise data) and Wikipedia, and a state library in Australia, unsurprisingly finding that the library catalogue is used much less frequently than the former two, but positing that the "fact that popular culture queries accounted for [a very] substantial proportion of Google and Wikipedia queries and almost no [library] catalogue queries indicates that, indeed, people do turn to different information resources for different subjects."
- **"Lexical clues" predict article quality**: A paper was presented at the 3rd Symposium on Web Society (SWS) last October[26] which sought to predict article quality based on eight different ratios derived from counting the number of sentences, words, diverse words, nouns, verbs, diverse nouns, diverse verbs and copulas in the article text. They trained a decision tree on a sample of 200 start-class and 200 featured articles (truncating each of the latter to 800 to 1000 words to arrive at a typical start-class article length) and then tested it on a different sample of 100 start-class and 100 featured articles, achieving precision and recall of more than 83% each.

References

[1] Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2011). Echoes of power: Language effects and power differences in social interaction. http://arxiv.org/abs/1112.3670 Open access

[2] http://www.technologyreview.com/blog/arxiv/27437/

[3] Soules, A. (2012). Where's the bio? Databases, Wikipedia, and the web. New Library World, 113(1/2), 77–89. Emerald Group Publishing Limited. DOI:10.1108/03074801211199068 Closed access

[4] https://sites.google.com/site/biographyanalysis/home

[5] http://lists.ala.org/wws/arc/les-l/2010-09/msg00002.html

[6] http://www.cs.rpi.edu/~magdon/courses/CASPfall2011.html

[7] Lavoie, A. (2011). Interaction vs. Homophily in Wikipedia Administrator Selection. http://assassin.cs.rpi.edu/~magdon/courses/casp/projects/Lavoie.pdf Open access

[8] Molnar, F. (2011). Link Prediction Analysis in the Wikipedia Collaboration Graph. http://assassin.cs.rpi.edu/~magdon/courses/casp/projects/Molnar.pdf Open access

[9] George, R. (2011). Link prediction on a Wikipedia dataset based on triadic closure. http://www.cs.rpi.edu/~magdon/courses/casp/projects/George.pdf Open access

[10] http://www.aeaweb.org/aea/2012conference/program/meetingpapers.php

[11] Zhu, Feng; Greenstein, S. (forthcoming). Is Wikipedia Biased? American Economic Review (Papers and Proceedings). http://www-bcf.usc.edu/~fzhu/wikipediabias.pdf Open access
[12] http://www.pbs.org/mediashift/2006/04/email-debatewales-discusses-political-bias-on-wikipedia111.html
[13] http://www.wikisym.org/2012/01/17/wikisym-2012-call-for-participation/
[14] http://wikimania2012.wikimedia.org/wiki/Submissions
[15] http://blog.wikimedia.de/2012/01/17/call-for-papers-wikipedia-academy-2012-research-and-free-knowledge/
[16] http://www.onlinecreation.info/?p=474
[17] http://lists.wikimedia.org/pipermail/wiki-research-l/2012-January/001799.html
[18] http://wikipapers.referata.com/
[19] Nielsen, F. A. (2011). Wikipedia research and tools: Review and comments. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6012 (working paper) Open access
[20] http://lists.wikimedia.org/pipermail/wiki-research-l/2012-January/001805.html
[21] Redondo-Olmedilla, J.-C. (2012). A Review of "Good Faith Collaboration: The Culture of Wikipedia." The Information Society, 28(1), 53–54. Routledge. http://dx.doi.org/10.1080/01972243.2011.632286 Closed access
[22] http://reagle.org/joseph/blog/social/wikipedia/gfc-web-cc-announce
[23] Colgrove, Caitlin; Neidert, Julia; Chakoumakos, R. (2011). Using Network Structure to Learn Category Classification in Wikipedia. http://www.stanford.edu/class/cs224w/proj/colgrove_Finalwriteup_v1.pdf Open access
[24] http://www.stanford.edu/class/cs224w
[25] Waller, V. (2011, November 8). Searching where for what: A comparison of use of the library catalogue, Google and Wikipedia. Library and Information Research. http://www.lirgjournal.org.uk/lir/ojs/index.php/lir/article/view/466 Open access
[26] Yanxiang, X., & Tiejian, L. (2011). Measuring article quality in Wikipedia: Lexical clue model. 2011 3rd Symposium on Web Society (pp. 141–146). IEEE. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6101286 Closed access

Issue 2(2): February 2012

CSCW 2012 in review; gender gap and conflict aversion; collaboration on breaking news; effects of leadership on participation; legacy of Public Policy Initiative

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

Wikipedia research at CSCW 2012

The annual 15th ACM conference on computer-supported cooperative work (**CSCW 2012**) featured two sessions about Wikipedia Studies. The first one was titled "Scaling our Everest [1]" (in amusing contrast to an earlier metaphor for the role of Wikipedia in that field of research: "the fruit fly of social software"), and covered four papers. A second session [2] likewise comprised four papers and notes. Below are some of the highlights from these two sessions.

Gender gap connected to conflict aversion and lower confidence among women

The Gender Gap hub on Meta.

Since January 2011, Wikipedia's "Gender gap" has received much attention from Wikimedians, researchers and the media – triggered by a *New York Times* article that cited the estimate that only 12.64% of Wikipedia contributors are female. That figure came from the 2010 UNU-MERIT study, which was based on the first global, general survey of Wikipedia users, conducted in 2008 with 176,192 respondents using a methodology that had raised some questions (e.g. about sample bias and selection bias), but other studies found similarly low ratios. A new paper titled "Conflict, Confidence, or Criticism: An Empirical Examination of the Gender Gap in Wikipedia"[3] has now delved further into the data of the UNU-MERIT study, examining the responses to questions such as "Why don't you contribute to Wikipedia?" and "Why did you stop contributing to Wikipedia?", finding strong support for the following three hypotheses:

- "**H1**: Female Wikipedia editors are less likely to contribute to Wikipedia due to the high level of conflict involved in the editing, debating, and defending process." ("Controlling for other factors females were 26% more likely to select 'I got into conflicts with other Wikipedia contributors' as a reason for no longer contributing. The coefficients for being afraid of being 'criticized' [31% higher probability to be selected by female users as a reason against becoming more active in Wikipedia], 'yelled at', and 'getting into trouble' are all significant".)
- "**H2**: Female Wikipedia editors are less likely to contribute to Wikipedia due to gender differences in confidence in expertise to contribute and lower confidence in the value of their contribution. "
- "**H3**: Female contributors are less likely to contribute to Wikipedia because they prefer to share and collaborate rather than delete and change other's work."

A fourth hypothesis likewise tested a conjecture that has been brought up several times in discussion about Wikipedia's gender gap:

- "**H4**: Female contributors are less likely to contribute to Wikipedia because they have less discretionary time available to spend contributing".

However, the paper's authors argued that this conjecture was not borne out by the data, instead finding that "men are 19% more likely to select 'I didn't have time to go on' as a reason for no longer contributing."

Making sense of NPOV

A paper titled "From Individual Minds to Social Structures: The Structuring of an Online Community as a Collective–Sensemaking Process" [4] looks at how Wikipedia editors talked about the Neutral point of view (NPOV) policy in the period of July 2005 to January 29, 2006, using Karl Weick's model of sensemaking and Anthony Giddens' theory of structuration for its theoretical approach. The paper's focus was on "how individual sensemaking efforts turn into interacts"; in other words, trying to understand how editors came to understand the NPOV policy through examining their posts. Editors' posts were differentiated into three types of questions (asking clarificatory questions, asking about behavior and the rules, and using questions as rhetorical devices) and answers (offering interpretation, explanation to others, and explanation to oneself).

Public Policy Initiative motivated students to become Wikipedians

Wikimedia Director Sue Gardner shows 11 packs of printer paper: the equivalent volume of content produced by students in the Wikipedia Public Policy Initiative.

In a paper titled "Classroom Wikipedia participation effects on future intentions to contribute"[5] (presentation slides [6]), five Michigan-based researchers looked at a sample of over 400 students who were involved in a pilot of the WMF education initiative (87% of whom were native speakers of English), and asked how likely the student-editors were to be become real editors after the end of their class projects, and what the relevant factors in such conversions are. They find that the student retention ratio is higher than the average editor retention ratio (while only 0.0002% of editors who make one edit become regulars, about 4% of students have made edits after their course ended). About 75% of the students preferred the Wikipedia assignment to a regular one, and major reasons for their enjoyment included the level of engagement in class, an appreciation of global visibility of the article, and the exposure to social media.

In related news, Erik Olin Wright, president of the American Sociological Association (ASA) who last year announced the organization's "Wikipedia Initiative", posted an overview[7] of a graduate seminar he conducted with a Wikipedia component. The students had to review a book, and use their newly gained knowledge to expand a relevant article on Wikipedia. In his assessment, Wright called the activity a "great success" and encouraged others to engage in similar activities.

High-tempo contributions: Who edits breaking news articles?

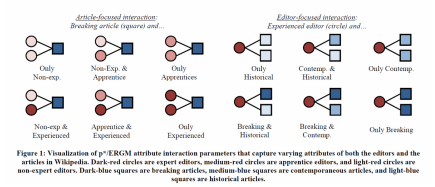


Figure 1: Visualization of p*/ERGM attribute interaction parameters that capture varying attributes of both the editors and the articles in Wikipedia. Dark-red circles are expert editors, medium-red circles are apprentice editors, and light-red circles are non-expert editors. Dark-blue squares are breaking articles, medium-blue squares are contemporaneous articles, and light-blue squares are historical articles.

Editor-focused and article-focused interactions emerging from a study of high-tempo collaboration in the English Wikipedia, Keegan et al. (2012).[8]

A team based at Northwestern University studied how topics of a specific nature find matching contributors in Wikipedia, or more precisely: "how editors with particular skills self-organize around articles requiring different forms of collaboration". The study[8] focused on the case of co-authorship in the context of breaking news articles. The authors note that such articles pose an interesting paradox: those that undergo a high-tempo editing cycle involving multiple contributors at once typically manifest quality issues, as the increased cost of interaction inhibits quality improvement work, yet in the unique case of breaking news articles, quality tends to remain very high despite multiple contributors attempting to make simultaneous edits with incomplete information or poor coordination.

The study uses revision data describing 58,500 contributions from 14,292 editors to 249 English Wikipedia articles about commercial airline disasters and represents them as a bipartite network characterized as article and editor nodes. A statistical model (p*/ERGM) is applied to estimate the likelihood of the creation of a link between a pair of nodes as a function of specific network properties or node attributes. The analysis focuses both on attributes of each set of nodes (e.g. whether an article is "breaking news", or the number of editor contributions) as well as properties of article-editor pairs as illustrated in the figure (*at right*). Some of the main results of the study were:

- Breaking news articles are more likely to attract editors.
- Breaking news articles are not more likely to get experts to work together: experienced editors work together on high tempo collaborations significantly less often than would be expected by chance.
- Experienced editors are unlikely to collaborate together on breaking news articles.

- Experienced editors tend to contribute to similar types of articles more than dissimilar types of articles (suggesting the existence of classes of experienced editors who mostly focus on breaking news topics).

How different kinds of leadership messages increase or decrease participation

Centralized or shared leadership?

Three social computing researchers from Carnegie Mellon University measured the "Effectiveness of Shared Leadership"[9] on the English Wikipedia – a model where leadership is not restricted to a few community members in a specialized role, but rather distributed among many. In an earlier paper (reviewed in a previous report), they had found evidence for shared leadership from an analysis of four million user talk page messages from a January 2008 dump of the English Wikipedia, classifying them (using machine learning) into four kinds of behavior indicating different kinds of "leadership": "transactional leadership" (positive feedback), "aversive leadership" (negative feedback), "directive leadership" (providing instructions) and "person-focused leadership" (indicated by "greeting words and smiley emoticons"). Based on this data, the present paper examines whether these four forms of messages increase or decrease the edit frequency of the user who receives them, also taking into account whether the message comes from an administrator or a non-administrator. Their first conclusion is that messages sent by both kinds of editors "significantly influenced other members' motivation", and secondly, they found that "transactional leaders and person-focused leaders were effective in motivating others, whereas aversive leaders' transactional and person-based leadership had the strongest effects, suggesting that interfaces and mechanisms that make it easier for editors to connect with, reward, and express their appreciation for each other may have the greatest benefits." (The sample predates the introduction of the "WikiLove" software extension which has exactly this goal.) Addressing a common objection by active Wikipedians in defense of warning messages, they acknowledge that "[p]eople may argue that reducing the activity of harmful editors is a positive impact of aversive leadership. However, considering the fact that there is much work to be accomplished in Wikipedia and the recent downward trend of active editors, pure aversive leadership should be avoided." The paper did not attempt to measure the quality of the work of the message recipients.

The researchers had to use a technique called propensity score matching to address the difficulty that true experimentation – for instance, separating users into control groups – was not possible in this purely observational approach. However, they separately examined the case of Betacommandbot, who had sent "more than half of the messages categorized as aversive leadership" in the sample, warning users who had uploaded a non-free image without a valid fair use rationale. Because these messages had been sent to editors regardless of whether their contributions were in violation of policy at the time they were made, "the Betacommandbot warning was a natural experiment, like a change in speeding laws, that was not induced by recipients' behavior". The effect of this warning was to decrease the recipients' edits by more than 10%.

Other CSCW 2012 contributions

- **Which edits get reverted?**: In "Learning from History: Predicting Reverted Work at the Word Level in Wikipedia"[10] researchers examined a sample of 150 articles from the English Wikipedia with over 1,000 revisions each. Every edit was classified according to whether it was a revert or not, and examined for these features: "the number of times each word is added or removed as two separate features, leading to feature spaces that are on average three to ten thousand words in size ... comment length, the anonymity of the editor, and his or her edit count and time registered on Wikipedia." The researchers then tried to construct a separate classifier for each article predicting whether a given edit would be reverted, with random decision tree forests turning out to be

most accurate, such that "the model .. obtained high accuracy [even] when vandalistic edits and bots were filtered out". Even when only taking the added words into account (ignoring user-based data and removed words), it was "still obtaining reasonable results". For the article genetic engineering, added words that made a revert likely were those "that violated policy or article conventions, had spelling errors, or had Wiki syntax errors", whereas use of terms specific to the article's subject made reverts less likely. As a possible application of their model, they speculate that it "could inform [new] editors when their edit is likely to be reverted, enabling them to reflect on and revise their contribution to increase its perceived value".

- **WikiProject's "Collaborations of the Week" help increase participation**: The authors of the above reviewed paper on shared leadership also presented a paper in the "Social Network Analysis" session, titled "Organizing without Formal Organization: Group Identification, Goal Setting and Social Modeling in Directing Online Production",[11] finding evidence for the effectiveness of "Collaboration of the Week (COTW)"-type article improvement drives on WikiProjects. (The *Signpost*'s "WikiProject report" series is cited at one point in the paper.)
- **Should a new wiki be "seeded" to invite participation?**: Apart from research specifically about Wikipedia, the conference featured many other results that are potentially of interest to Wikimedians and Wikipedia researchers. For example, a paper titled "Bootstrapping wikis: Developing critical mass in a fledgling community by seeding content" [12] reported on an experiment with 96 students who were asked to spend 20 minutes on contributing to a new MediaWiki-based course wiki, and "found that users tend to contribute more content, and more unstructured content, when they are given a blank slate. This suggests that bootstrapping is not always a positive. However, users tend to contribute content roughly similar to any seeded content. Bootstrapping can be used to direct user effort toward contributing speci?c types of content".
- Two other papers presented at CSCW 2012 focused on the **editing behavior of new Wikipedians**[13] and on collaboration in **breaking news articles**.[14]

Wikipedia discourse on Europe analyzed

A master thesis by Dušan Miletic on *Europe According to English Wikipedia: Open-sourcing the Discourse on Europe*[15] looks at the nature of the discourse on Europe in the English Wikipedia, employing Foucauldian discourse analysis, which focuses on analyzing the power in relationships as expressed through language. The article notes that "changes to the statements defining what Europe is, which hold the cardinal role in the discourse, had much more significance than others." In other words, the editors who succeeded in changing the definition of Europe were subsequently able to have their points of view better represented in the remainder of the article. Another finding suggests that the definition of European culture was much more difficult to arrive at, and spawned many more revisions throughout the article, than the discussion of the geography of Europe. Another aspect discussed in the article is the blurry boundary between Europe and the European Union. The article concludes that the borders of European culture are not the same as the borders of geographical Europe, and hence, that the difficult task of defining Europe – and revising the Wikipedia article – is bound to continue.

The significance of the first edit

A paper titled "Enrolled Since the Beginning: Assessing Wikipedia Contributors' Behavior by Their First Contribution"[16] by researchers at Telecom Bretagne looks at an editor's first contribution as an indicator of her future level of involvement in the project. After having discovered Wikipedia, the sooner one makes their first edit, the higher the likelihood they will continue editing. Reasons for the first edit matter, as those who just want to see how a wiki works are less likely to keep editing than those who want to share (improve) something specific, content-wise. Making a minor edit is much less likely to result in a highly active editor; those who will become very active are often those whose very first edit required a large investment of time. As the authors note, "it seems that those who will become the core editors of the community have a clearly defined purpose since the beginning of their participation and don't waste their time with minor improvements on existing articles". Finally, the authors find that

having a real life contact who shows one how to edit Wikipedia is much more likely to result in that person becoming a regular Wikipedia contributor, compared to people who learn how to edit by themselves.

Given enough eyeballs, do articles become neutral?

Building on their previously reviewed research, Greenstein and Zhu ask[17] "will enough eyeballs eliminate or decrease the amount of bias when information is controversial, subjective, and unverifiable?" Their research calls this into question, by taking a statistical approach to measuring bias in Wikipedia articles about US political topics, which uses Linus' Law ("Given enough eyeballs, all bugs are shallow") as a null hypothesis.

They rely on a slant index previously developed for studying news media bias, which specifies certain code words as indicating Republican or Democratic bias. Within their sample of 28,382 articles relating to American politics, they find that the category and vintage of an article are most predictive of bias. "Topics of articles with the most Democrat words are civil rights, gun control, and homeland security. Those with the most Republican words are abortion, foreign policy, trade, tax reform, and taxation. ... [T]he slant and bias are most pronounced for articles born in 2002 and 2003". While they do not find a neutral point of view within each article or topic, across articles, Wikipedia balances Democratic and Republican points of view.

Yet answering "Why did Wikipedia become less biased over time?" is more challenging. They classify explanatory variables into three groups: attention and editing; dispersion of contributions; and article features. The narrow interpretation of Linus' Law would make attention and editing the only relevant feature (not supported by their data), while a broader interpretation would also take dispersion into account (weak support from their data). While both the number of revisions and the number of editor usernames are statistically significant, they work in opposite directions. Pageviews, while also statistically significant, are unavailable before February 2007. They also suggest questions for further work, including improvements to their revision sampling (they "divide [each article's] revisions into ten revisions of equal length") and overall sampling method (which uses the same techniques as their earlier work).

Navigating conceptual maps of Wikipedia language editions

A paper from this year's Conference on Human Factors in Computing Systems (CHI 2012) entitled "Omnipedia: Bridging the Wikipedia Language Gap"[18] presents the features of *Omnipedia*, a system that enables readers to analyse up to 25 language editions of Wikipedia simultaneously. The study also includes a review of the challenges that the architects faced in building the Omnipedia system, as well as the results of initial user testing. According to the authors, language barriers produce a silo effect across the encyclopedias, preventing users from being able to access content unique to different language editions. Omnipedia, they write, reduces the silo effect by enabling users to navigate different concepts (over 7.5 million of them) from up to 25 language editions of Wikipedia, highlighting similarities and differences in an interactive visualization that shows which concepts different editions mention and how each of those topics is discussed.

The authors provide the example of the English Wikipedia article on conspiracy theory, showing how it discusses many topics – from "Moon landing" to "Kennedy assassination". Other language editions contain articles on the same concept, including *Verschwörungstheorie* in the German Wikipedia and *teoria conspirativa* in the Spanish Wikipedia. Omnipedia consolidates these articles into a single "multilingual article" on conspiracy theories, showing which language editions have topics discussed in only one language edition and which have those discussed in multiple language editions.

The paper concludes with the results of user testing, showing how the volume of single-language topics was "a revelation to the majority of users" but also how users targeting concepts they thought might reveal differences in perspective (for example on "Climate scepticism" or the "War on the Terror") actually had fewer differences than anticipated. The authors conclude by highlighting their contributions to this area of study, including a system that for the first time allows simultaneous access to large numbers of Wikipedia language editions – powered by several new

algorithms that they assert "preserve diversity while solving large-scale data processing issues" – and a demonstration of the value of Omnipedia to user analysis of concepts explored in different language editions.

Briefly

- **Taxonomy extraction in Wikipedia**. Mike Chen recently defended an MSc thesis at Ohio University focused on extracting a taxonomy from Wikipedia articles.[19] A similar problem is discussed in a paper by a team based at Technische Universität Darmstadt [20] which presents a solution tackling two of the main challenges for building a robust category system for Wikipedia: multilingualism and the sparse connectivity of the semantic network (i.e. the fact that users do not identify resources on the same topic with identical tags). A third paper by researchers at Telecom Bretagne[21] develops a method to extract a tree from the Wikipedia category graph and tests its classification precision against a corpus from Wikinews.
- **Dynamics of Wikipedia conflicts**. A team of researchers from Hungary studied the dynamics of controversies in Wikipedia and analyzed their temporal characteristics.[22] They find a correspondence between conflict and burstiness of activity patterns and identify patterns that match cases eventually leading to consensus as opposed to articles where a compromise is far from achievable.
- **Do social norms influence participation in Wikipedia?**. A paper titled "Factors influencing intention to upload content on Wikipedia in South Korea: The effects of social norms and individual differences"[23] reported on the results of a survey, analyzing responses by 343 South Korean students ("uploading" meaning any form of contributing to Wikipedia). Among the findings was that "users of Wikipedia presented higher perceived injunctive norm and greater self-efficacy [roughly: confidence in their ability] toward the intention to upload content on Wikipedia than non-users. These findings can be understood as follows: uploading content on Wikipedia is a socially desirable act given that it contributes to knowledge sharing, and thus, for the people who already use Wikipedia, they might feel that they are urged by their social groups to upload content on the site as a way of participating in the making of collective intelligence."
- **Link disambiguation and article recommendations**. An MSc thesis in computer science defended by Alan B Skaggs from the University of Maryland proposes a statistical topic model to suggest new link targets for ambiguous links in Wikipedia articles.[24] Three Stanford University students in computer science proposed a recommendation engine for Wikipedia articles using only a small set of articles *liked* by a population of users as training data.[25]

References

[1] http://www.cscw2012.org/program/papersnotes.php#31

[2] http://www.cscw2012.org/program/papersnotes.php#36

[3] Collier, B., & Bear, J. (2012). Conflict, criticism, or confidence. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12 (p. 383). New York, New York, USA: ACM Press. **PDF** (http://dl.acm.org/citation.cfm?id=2145204.2145265) • **DOI** (http://dx.doi.org/10.1145/2145204.2145265)Closed access

[4] Nagar, Y. (2012) What do you think?: the structuring of an online community as a collective-sensemaking process. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12. New York, New York, USA: ACM Press. **PDF** (http://mit.edu/ynagar/www/papers/Nagar_Y_Collective_Sensemaking.CSCW_2012.pdf) **DOI** (http://dx.doi.org/10.1145/2145204.2145266) Open access

[5] Zube, P., Velasquez, A., Ozkaya, E., Lampe, C., & Obar, J. (2012). Classroom Wikipedia participation effects on future intentions to contribute. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12 (p. 403). New York, New York, USA: ACM Press. **PDF** (http://dl.acm.org/citation.cfm?id=2145204.2145267) **DOI** (http://dx.doi.org/10.1145/2145204.2145267) Closed access

[6] http://www.slideshare.net/clifflampe/cscw-2012-note-wikipedia-public-policy-initiative

[7] Erik Olin Wright (2012) Writing Wikipedia Articles as a Classroom Assignment, *ASA Newsletter* (Teaching Sociology), February 2012 **PDF** (http://www.ssc.wisc.edu/~wright/ASA/Writing Wikipedia Articles as a Classroom Assignment.pdf) Open access

[8] Keegan, Brian, Darren Gergle, and Noshir Contractor (2012). Do Editors or Articles Drive Collaboration? Multilevel Statistical Network Analysis of Wikipedia Coauthorship. In *2012 ACM Conference on Computer Supported Cooperative Work (CSCW '12)*. **PDF** (http://www.brianckeegan.com/papers/CSCW12.pdf) Open access

[9] Zhu, H., Kraut, R., & Kittur, A. (2012). Effectiveness of shared leadership in online communities. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12 (p. 407). New York, New York, USA: ACM Press. **PDF** (http://www.cs.cmu.edu/~haiyiz/ttt/SharedLeadership_fifthpdf.pdf) • **DOI** (http://dx.doi.org/10.1145/2145204.2145269) Open access

[10] Rzeszotarski, J., & Kittur, A. (2012). Learning from history. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work – CSCW '12* (p. 437). New York, New York, USA: ACM Press. **DOI** Closed access

[11] Zhu, H., Kraut, R., & Kittur, A. (2012). Organizing without Formal Organization: Group Identification, Goal Setting and Social Modeling in Directing Online Production. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work – CSCW '12* (p. 935). New York, New York, USA: ACM Press. **PDF** (http://www.cs.cmu.edu/~haiyiz/papers/COTW.pdf) • **DOI** Open access

[12] Solomon, J., & Wash, R. (2012). Bootstrapping wikis. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12 (p. 261). New York, New York, USA: ACM Press. **PDF** (http://www.rickwash.com/papers/bootstrapping-cscw.pdf) • **DOI**Open access

[13] Antin, J., Cheshire, C., & Nov, O. (2012). Technology-mediated contributions. Editing Behaviors Among New Wikipedians. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* - CSCW '12 (p. 373). New York, New York, USA: ACM Press. **PDF** (http://faculty.poly.edu/~onov/Antin_Chehsire_Nov_WPP_CSCW_2012.pdf) • **DOI** Open access

[14] Keegan, Brian C (2012). Breaking news on Wikipedia: Dynamics, structures, and roles in high-tempo collaboration. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion – CSCW '12*, New York, New York, USA: ACM Press, 2012. **PDF**Closed access

[15] Miletic, Dušan (2012). Europe According to English Wikipedia. Open-sourcing the Discourse on Europe, Masters Thesis, Jagiellonian University **PDF** (http://scripties.let.eldoc.ub.rug.nl/FILES/root/Master/DoorstroomMasters/Euroculture/2012/d.miletic/MA-2145189-D_Miletic.pdf)Open access

[16] Dejean, Sylvain, and Nicolas Jullien (2012). Enrolled Since the Beginning: Assessing Wikipedia Contributors' Behavior by Their First Contribution. *SSRN Electronic Journal* **PDF** (http://ssrn.com/paper=1980806)Open access

[17] Zhu, Feng, and Shane Greenstein (2012). Collective Intelligence and Neutral Point of View: The Case of Wikipedia **PDF** (http://www.colorado.edu/econ/seminars/greenstein.pdf)Open access

[18] Bao, Patti, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle (2012) Omnipedia: Bridging the Wikipedia Language Gap. In: *Proc. CHI 2012*. **PDF** (http://www.brenthecht.com/papers/bhecht_CHI2012_omnipedia.pdf)Open access

[19] Chen, Mike (2011). Taxonomy Extraction from Wikipedia, Masters Thesis, Ohio University. **PDF** (http://etd.ohiolink.edu/view.cgi/Chen Mike.pdf?ohiou1320342712)Open access

[20] Garcìa, Renato Dominguez, Philipp Scholl, and Christoph Rensing (2011). Supporting Resource-based Learning on the Web using automatically extracted Large-scale Taxonomies from multiple Wikipedia versions. *Advances in Web-based learning - ICWL 2011*, LNCS 7048. **PDF** (ftp://www.kom.tu-darmstadt.de/papers/DSR11.pdf)Open access

[21] Haralambous, Yannis, and Vitaly Klyuev (2012). Wikipedia Arborification and Stratified Explicit Semantic Analysis, *Computation and Language* (January 30, 2012): 13. **PDF** (http://arxiv.org/pdf/1201.6224.pdf) Open access

[22] Yasseri, Taha;Sumi, Robert;Rung, András;Kornai, András;Kertész, János (2012) Dynamics of conflicts in Wikipedia. *Physics and Society; Data Analysis, Statistics and Probability*. ArXiV (February 16, 2012). **PDF** (http://arxiv.org/abs/1202.3643)Open access

[23] Park, Namkee, Hyun Sook Oh, and Naewon Kang (2012). Factors influencing intention to upload content on Wikipedia in South Korea: The effects of social norms and individual differences. *Computers in Human Behavior* 28(3), May 2012: 898-905. **DOI** (http://dx.doi.org/10.1016/j.chb.2011.12.010) Closed access

[24] Skaggs, Bradley Alan (2011). *Topic Modeling for Wikipedia Link Disambiguation*, Masters Thesis, University of Maryland **HTML** (http://hdl.handle.net/1903/12383)Open access

[25] Rothfels, John, Brennan Saeta, and Emin Topalovic (2011). *A recommendation engine for Wikipedia articles based on constrained training data*. **PDF** (http://cs229.stanford.edu/proj2011/TopalovicRothfelsSaeta-ARecommendationEngineForWikipediaArticlesBasedOnConstrainedTrainingData.pdf)Open access

Issue 2(3): March 2012

Predicting admin elections by editor status and similarity; flagged revision debates in multiple languages; Wikipedia literature reviewed

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

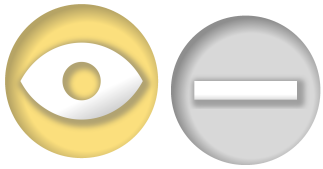
How editors evaluate each other: effects of status and similarity

A team of social computing researchers based at Stanford and Cornell University studied how users evaluate each other in social media.[1] Their paper, presented at the 5th ACM Web Search and Data Mining Conference [2] (WSDM '12), focuses on three main case studies: Wikipedia, StackOverflow and Epinions. User-to-user evaluations, the authors note, are jointly influenced by the properties of the evaluator and the target; as a result, differences in properties between the target and the evaluator should be expected to affect the evaluation. The study looks specifically at how differences in topic expertise and status affect peer evaluations. The Wikipedia case focuses on requests for adminship (RfAs), the most prominent example of peer evaluation in Wikipedia and a topic that has attracted considerable attention in the literature (*Signpost* research coverage: September 2011, October 2011, January 2012). Similarity is measured based on article co-authorship, and status as a function of an editor's number of contributions. Previous research by the same authors showed that the probability an evaluator will evaluate a target user positively drops dramatically when the status of the two users is very similar, and there is general evidence that homophily and similarity in editing activity have a strong influence on peer evaluation in RfAs. The study identifies two effects that jointly account for this singular finding:

- “Elite” or high-status users are more likely to participate in evaluations about other users who are active in their areas of interest or expertise.
- Low-status users tend to be judged differently than those with moderate or high status

In a direct application of these results, dubbed *ballot-blind prediction*, the authors show how the outcome of an RfA can be accurately predicted by a model that simply considers the first few participants in a discussion and their attributes, without looking at their actual evaluations of the target.

Sociological analysis of debates about flagged revisions in the English, German and French Wikipedias



At the center of debates on "Coercion or empowerment": Icons signifying accepted (left) and not yet accepted (right) revisions under a flagged revisions scheme

In an article to appear in *Ethics and Information Technology*, Paul B. de Laat analysed debates occurring in the English, German and French Wikipedias about the evolution of the rules governing new edits.[3] As noted by the analysis of the English Wikipedia's rules, by Butler et al., 2008,[4] these rules are numerous and have increased in number and complexity; they range from the more formal and explicit (intellectual property rights) to the more informal.

De Laat's work is based on a study of the discussions around the proposal to introduce a system of reviewing edits before they appear on screen (flagged revisions, discussed on English Wikipedia at Wikipedia:Flagged revisions). It focuses on the perennial debate around the construction of knowledge commons theorized by Elinor Ostrom:[5] being a collective, open project, it must be accessible to most, but as its production becomes important for its "owners" (readers and producers), boundaries have to be set to protect its integrity. De Laat's article describes and analyzes the

tensions and permanent adjustments needed to manage these apparently opposed goals.

In a Weberian analysis of bureaucracy, applicable to Wikipedia policies, he shows that two views can be invoked to explain the intensity of the discussions. He summarizes the debate as a clash between (i) those who saw the flagged revisions as "a useful tool for curbing vandalism", enabling and empowering users and editors, and (ii) those who denounced it as "a superfluous bureaucratic device that violates egalitarian principles of participation", designed to introduce a more controlled and hierarchical environment. He muses that "an intriguing question that remains to be answered, of course, is: What brought the three language communities to ultimately choose or reject such a review system? Why is it that, each in their own ways, the Germans voted for acceptance, the French for rejection, while the English have been wavering all the time between acceptance and rejection"? (p. 11) This question, and Wikipedians' views of flagged revisions, can shine light onto what kind of community Wikipedia should be, according to various factions of editors. As De Laat answers it, "many of those who reject the system of review do so from a vision of Wikipedia as an unbounded community that shares knowledge without mutual control and suspicion, while many of those who embrace the review system do so because they have a vision of Wikipedia as an organization producing reliable knowledge that keeps vandalism outside its borders". De Laat suggests that further research is needed to fully understand the factors affecting the decisions on different Wikipedias taken with regard to flagged revisions, postulating a hypothesis to be tested in further research that "those whose mother tongue is German may possibly be more deferential to hierarchy than those who speak either French or English, and therefore may prefer the order and respectability introduced by a system of reviewing".

Understanding collaboration-related dialog in Simple English Wikipedia

In a paper published by the European chapter of the ,[6] Oliver Ferschke and coauthors describe a study of talkpages on the Simple English Wikipedia. This paper uses and as a theoretical framework for studying how authors use discussion pages to collaborate on article improvement. They have released a freely downloadable corpus of 100 segmented and annotated talk pages [7], called the Simple English Wikipedia Discussion Corpus [7] and based on a new annotation schema for coordination-related dialog acts. Their schema uses 17 categories, grouped into these four top-level categories: article criticism, explicit performative announce, information content, and the interpersonal. The authors use their corpus to develop a machine-learning-based pipeline for dialog act classification, which they describe but which is not freely available. They provide a useful discussion of and good pointers to seminal and new research in dialog acts. (A longer, editable summary [8] is available on AcaWiki.)

Majority of UK academics prohibit students from using Wikipedia, but use it just as frequently themselves

An article appearing in "Teaching in Higher Education"[9] "discusses the use of Wikipedia by academics and students for learning and teaching activities at Liverpool Hope University, [considering] the findings to be indicative of Wikipedia use at other British universities". Having sent email invitations to all staff and students at the university, they received responses from 133 academics and 1222 students. 75% of the student respondents said they used Wikipedia for "some purpose", which according to the authors indicates that Wikipedia use "has risen appreciably in a short period of time" among British university students, citing a 2009 study[10] which had put that number at only 17.1%. "However", they cautioned, usage was "significantly lower than usage in the USA."

Among the surveyed teaching staff, almost the same percentage (74%) used Wikipedia "for some purpose" as their students—but just 24% of them "tell their students to use Wikipedia for Learning and Teaching purposes, with 18% having not mentioned it to students and 58% having expressly told them not to." The independence of academics' answers to these two questions is highlighted by the authors as

> "a key finding of the survey: there is little difference between academics that permit their students to use Wikipedia and those who do not in respect to their own use. In particular, amongst both groups, academics that used Wikipedia 'frequently' seem to exhibit similar usage profiles. It was indicated in the commentary that the

> critical difference was that they have the scholarly expertise to determine what material on Wikipedia was 'correct' and that which was not."

In the conclusion the authors observe that "a significant proportion of what we would see as enlightened academics at Liverpool Hope and no doubt elsewhere realise that it is pointless to try to hold back the online tide of Wikipedia. Instead, they try to give guidance in the way that students consult it: for clarification, references, comparison and definitions."

A systematic review of the Wikipedia literature

"The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia"[11] is the title of a working paper which promises to be a major milestone in Wikipedia research. It is an attempt to synthesize a broad-based literature review of scholarly research on Wikipedia. The task of creating a comprehensive database of such publications has seen several efforts before and its difficulties were explored in a well-attended workshop at last year's WikiSym conference (see the October issue of this research report).

The authors intend to release their findings in a "Web 2.0" format through their wiki [12] by the end of May 2012. The current paper is impressive in scope, but at 71 pages badly in need of a table of contents (the current version does not seem to adhere to any consistent manual of style, with headings using different font sizes and even colors) and clarifications (the current distinction between findings on p.12 and discussion on p. 19 seems somewhat arbitrary; the authors at one point promise a discussion of over 2,000 articles and in other places talk of a sample of 139) – perhaps due to its genesis (see below). Keeping in mind this is just a draft paper, we hope the final paper will have an improved flow and transparency. The presented methodology is useful for those interested in learning how to analyze large, thematic bodies of work using online databases. In one of their major contributions, the authors intend to present an overview of Wikipedia research grouped by themes (keywords), such as for example discussing research done on "vandalism reversion", "thesaurus construction" or "attitude towards Wikipedia". While the current draft is not yet comprehensive, it shows much potential, and in practice their wiki, which already groups the content with categories, may prove more useful as a reference work.

As explained [13] by one of the authors, the paper merges two existing efforts, both of which already published drafts last year. And by choosing wikilit.referata.com as their platform, they embrace the work of a third party, Wikimedian User:Emijrp's "Wikipapers" [14] wiki on the same domain. This follows discussions between the three parties reported in the January issue of this research report ("New effort at comprehensive wiki research literature database"). On the wiki, the authors acknowledge [15] the modest efforts of a fourth party, namely this research report (which just released [16] a dataset of all publications covered until the end of 2011): "We do not include any items published after June 2011, after which the Wikimedia Research Newsletter was formally inaugurated; we're letting them pick up from where we stop."

Briefly

- **External links in the English Wikipedia**: A short paper by a team of Greek researchers presents statistics on the nature and quality of external links in the English Wikipedia, based on the October 2009 XML dump.[17] The analysis, although based on an outdated dataset, reveals insights into the distribution of links per article, the relation between external links and article length, and the proportion of dead links, which they quantify as 18% of the links in their corpus.
- **Wikipedia articles on StumbleUpon**: Two short papers which are to be presented at a workshop titled "Searching 4 Fun!" [18] collocated with the upcoming European Conference on Information Retrieval concern Wikipedia: "Serendipitous Browsing: Stumbling through Wikipedia"[19] examines which Wikipedia articles are being featured by users of the social bookmarking site StumbleUpon. Based on a sample consisting of a random selection of half of the articles from the October 2011 dump, 15.13% of the articles of the English Wikipedia are contained in StumbleUpon's index (as opposed to less than 1% of both the French and the German Wikipedia,

according to an initial investigation). The 100 articles with the most views by StumbleUpon users contained only one featured article, but twelve lists – among them the number one, the list of unusual deaths, which belongs to the "Bizarre/Oddities" category on StumbleUpon, as do four of the other top ten articles. A second w:position paper is titled "Searching Wikipedia: Learning the Why, the How, and the Role Played by Emotion"[20] proposes to examine users' search behavior, employing diary studies and a custom-built Firefox extension asking Wikipedia readers to record details about their informational requirement and the motivating situation driving the search.

- **Citations of open access articles in Wikipedia**: An ArXiv preprint by researchers based at UNC-Chapel Hill and the National Evolutionary Synthesis Center, studying "indicators of scholarly impact in social media" (or "altmetrics"), reports on the number of citations to the open access scholarly literature that can be found in Wikipedia.[21] The study suggests that 5% of all 24,331 articles ever published until November 2010 in the seven journals of open-access publisher Public Library of Science (PLoS) are cited in Wikipedia. More statistics on the number of scholarly citations in Wikipedia by language and by publisher are available via the Wikipedia cite-o-meter [22].
- **First results from Article Feedback v5**: The Wikimedia Foundation reported new results from the first stage of experiments with a fully redesigned article feedback tool.[23] The full report indicates that 45% of all reader suggestions (sampled from a random list of AFT5-enabled articles) were considered useful via a blind assessment performed by Wikipedia editors. The report also identifies differences in the overall volume of feedback posted via different designs and finds that asking readers to suggest what they were looking for outperforms comments with ratings.
- **Augmenting Wikipedia articles with Europeana items**: A paper titled "Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia"[24] proposes a mechanism that allows users "to browse Wikipedia articles, which are augmented with items from the cultural heritage collection. Using Europeana as a case-study we demonstrate the effectiveness of our approach for encouraging users to spend longer exploring items in Europeana compared with the existing search provision."
- **Semantics for genes**: Biochemists from the Gene Wiki project on Wikipedia report on "Building a biomedical semantic network in Wikipedia with Semantic Wiki Links".[25] Among other things, the paper mentions the introduction of Template:SWL, an attempt to emulate some aspects of Semantic MediaWiki using Wikipedia's existing (non-semantic) MediaWiki version.
- **Live version of DBpedia**: A paper[26] by the team behind the DBpedia project (which extracts structured data from Wikipedia) promises to explain the techniques behind a recent improvement, avoiding lags caused by infrequent updates. According to the abstract, "Wikipedia provides DBpedia with a continuous stream of updates, i.e. a stream of recently updated articles. DBpedia-Live processes that stream on the fly to obtain RDF data and stores the extracted data back to DBpedia."

References

[1] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Effects of user similarity in social media. *Proceedings of the fifth ACM international conference on Web search and data mining* - WSDM '12(p. 703). New York, New York, USA: ACM Press. **DOI** (http://dx.doi.org/10.1145/2124295.2124378) • **PDF** (http://www-cs.stanford.edu/people/ashton/pubs/wsdm-sim.pdf) Open access

[2] http://wsdm2012.org/

[3] de Laat, P. B. (2012). Coercion or empowerment? Moderation of content in Wikipedia as 'essentially contested' bureaucratic rules. *Ethics and Information Technology*, 1–13. Springer Netherlands. **DOI** (http://dx.doi.org/10.1007/s10676-012-9289-7) Open access

[4] Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems* – CHI '08 (p. 1101). New York, New York, USA: ACM Press. **DOI** (http://dx.doi.org/10.1145/1357054.1357227) • **PDF** (http://hci.uma.pt/courses/socialweb08F/5/butler.pdf) Open access

[5] Hess, Charlotte and Ostrom, Elinor (2006) A Framework for Analyzing the Knowledge Commons, in Hess, C., & Ostrom, E. (Eds.). *Understanding Knowledge as a Commons: From Theory to Practice*. MIT Press, 2006, pp. 41–81 Closed access

[6] Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2012). **PDF** (http://www.ukp.

tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/EACL_2012_OF.pdf) Open access

[7] http://www.ukp.tu-darmstadt.de/data/wikidiscourse

[8] http://acawiki.org/Behind_the_article:_Recognizing_dialog_acts_in_Wikipedia_talk_pages

[9] Knight, C., & Pryke, S. (2012). Wikipedia and the University, a case study. *Teaching in Higher Education*, 1–11. Routledge. **DOI** (http://dx.doi.org/10.1080/13562517.2012.666734) Closed access

[10] Hampton-Reeves, S., Mashiter, C., Westaway, J., Lumsden, P., Day, H., Hewertson, H., & Hart, A. (2009). *Students' Use of Research Content in Teaching and Learning Behaviour*. JISC, **PDF** (http://www.jisc.ac.uk/media/documents/aboutus/workinggroups/studentsuseresearchcontent.pdf) Open access

[11] Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2012). The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. *SSRN eLibrary*. SSRN. **HTML** (http://ssrn.com/paper=2021326) Open access

[12] http://wikilit.referata.com/wiki/Main_Page

[13] http://lists.wikimedia.org/pipermail/wiki-research-l/2012-March/001871.html

[14] http://wikipapers.referata.com/wiki/Main_Page

[15] http://wikilit.referata.com/w/index.php?title=Main_Page&oldid=2652#Key_details_about_this_literature_review

[16] https://blog.wikimedia.org/2012/03/16/wikimedia-research-newsletter-first-volume-new-features/

[17] Tzekou, P., Stamou, S., Kirtsis, N., & Zotos, N. (2011). "Quality assessment of Wikipedia external links." In J. Cordeiro & J. Filipe (Eds.), WEBIST 2011, *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (pp. 248-254). **PDF** (http://www.dblab.upatras.gr/download/nlp/NLP-Group-Pubs/11-WEBIST_Wikipedia_External_Links.pdf) Open access

[18] http://ceur-ws.org/Vol-836/

[19] Hauff, C., & Houben, G.-J. (2012). Serendipitous Browsing: Stumbling through Wikipedia. In D. Elsweiler, M. L. Wilson, & M. Harvey (Eds.), *Proceedings of the "Searching 4 Fun!" workshop, collocated with the annual European Conference on Information Retrieval* (ECIR2012) Barcelona, Spain, April 1, 2012. (pp. 21-24) **PDF** (http://ceur-ws.org/Vol-836/paper7.pdf) Open access

[20] Knäusl, H. (2012). Searching Wikipedia: Learning the Why, the How, and the Role Played by Emotion. In D. Elsweiler, M. L. Wilson, & M. Harvey (Eds.), *Proceedings of the "Searching 4 Fun!" Workshop, collocated with the annual European Conference on Information Retrieval* (ECIR2012) Barcelona, Spain, April 1, 2012. (pp. 14-15). **PDF** (http://ceur-ws.org/Vol-836/paper5.pdf) Open access

[21] Priem, J., Piwowar, H. A., & Hemminger, B. H. (2012). Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact. *ArXiV*. **PDF** (http://arxiv.org/abs/1203.4745)Open access

[22] http://toolserver.org/~dartar/cite-o-meter/

[23] Florin, F., Fung, H., Halfaker, A., Keyes, O., & Taraborelli, D. (2012). Helping readers improve Wikipedia: First results from Article Feedback v5. *Wikimedia Foundation blog*. **HTML** (http://blog.wikimedia.org/2012/03/21/helping-readers-improve-wikipedia-first-results-from-article-feedback-v5/) Open access

[24] Hall, M. M., Clough, P. D., Lopez de Lacalle, O., Soroa, A., & Agirre, E. (2012). Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia. **PDF** (http://work.room3b.eu/wp-content/papercite-data/pdf/halletal2012a.pdf) Open access

[25] Good, B. M., Clarke, E. L., Loguercio, S., & Su, A. I. (2012). Building a biomedical semantic network in Wikipedia with Semantic Wiki Links. *Database : The Journal of Biological Databases and Curation*, 2012, **DOI** (http://dx.doi.org/10.1093/database/bar060)Open access

[26] Morsey, M., Lehmann, J., Auer, S., Stadler, C., & Hellmann, S. (2012). DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: Electronic library and information systems*, 46(2), 2. Emerald Group Publishing Limited. **PDF** (http://www.emeraldinsight.com/journals.htm?articleid=17020943&show=abstract) Closed access

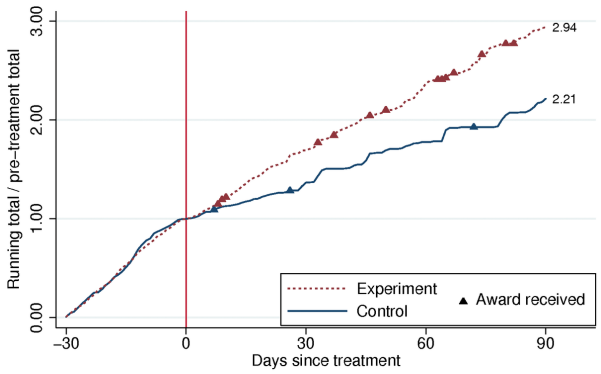
Issue 2(4): April 2012

Barnstars work; Wiktionary assessed; cleanup tags counted; finding expert admins; discussion peaks; Wikipedia citations in academic publications; and more

With contributions by: Lambiam, Piotrus, Jodi.a.schneider, Amir E. Aharoni, DarTar, Tbayer, Steven Walling, Junkie.dolphin and Protonk

Recognition may sustain user participation

To gain insight in what makes Wikipedia tick, two researchers from the Sociology Department at Stony Brook University conducted an experiment with barnstars.[1] They were surprised by what they found.



The relative number of edits by Wikipedians who had randomly received barnstars (red) and by the control group whose members hadn't (blue).

Professor Arnout van de Rijt and graduate student Michael Restivo wanted to test the hypothesis according to which *receiving recognition for one's work in an informal peer-based environment such as Wikipedia has a positive effect on productivity*. To test their hypothesis, they determined the top 1% most productive English Wikipedia users among the currently active editors who had yet to receive their first barnstar. From that group they took a random sample of 200 users. Then they randomly split the sample into an experimental group and a control group, each consisting of 100 users. They awarded a barnstar to each user in the experimental group; the users in the control group were not given a barnstar. The researchers found their hypothesis confirmed: the productivity of the users in the experimental group was significantly higher than that of the control group. What really took the researchers by surprise was how long-lasting the effect was. They followed the two groups for 90 days, observing that the increase in contribution level for the group of barnstar recipients persisted, almost unabated, for the full observation period.

One major factor the experiment did not take into account was whether it mattered who delivered barnstars and whether they were anonymous, registered, or known members of the Wikipedia community. During the experiment, it was noted on the Administrator's noticeboard/Incidents page that a seemingly random IP editor was "handing out barnstars", which led to some suspicion from Wikipedians. The thread was closed after User:Mike Restivo confirmed he accidentally logged out when delivering the barnstars. He did not, however, declare his status as a researcher, and the group's paper does not disclose that the behavior was considered unusual enough to warrant such a discussion thread.

Can Wiktionary rival traditional lexicons?

Wiktionary received an extensive assessment[2] as a potential rival to expert-built lexicons.

A chapter titled "Wiktionary: a new rival for expert-built lexicons?"[2] in a collection on electronic lexicography to appear with Oxford University Press contains a description and critical assessment of Wikipedia's second oldest sister project (which will celebrate its 10th anniversary in December this year) – subtitled "Exploring the possibilities of collaborative lexicography", which it calls a "fundamentally new paradigm for compiling lexicons".

The article describes in detail the technical and community features of Wiktionary. Though it is not immediately clear, the article's focus is on several language editions and not just English (as often happens in research about Wikipedia and its sister projects). The article gives a comprehensive account of the coverage of the world's languages by the various Wiktionary language editions. There is a critical analysis of Wiktionary's content, first with what appears to be a thorough statistical comparison with other dictionaries and wordnets, including an examination of the overlaps in the lexemes covered, which the authors found to be surprisingly small.

Number of native terms (p.17)	Wiktionary	wordnets	Roget's Thesaurus	OpenThesaurus
English language	352,865	148,730 (WordNet)	59,391	
German language	83,399	85,211 (GermaNet)		58,208
Russian language	133,435	130,062 (Russian WordNet)		

The article notes an important characteristic of open wiki projects: they allow "updating of the lexicons immediately, without being restricted to certain release cycles as is the case for expert-built lexicons" (p. 18). Though this characteristic is obvious to experienced Wikimedians, it is frequently overlooked. The discussion of the organization of polysemy and homonymy is comprehensive, although limited to the English Wiktionary. Other language editions may do it differently. The article notes that "it is a serious problem to distinguish well-crafted entries from those that need substantial revision by the community", which is good constructive criticism. The paragraphs about "sense ordering" make some vague claims (e.g. "Although there is no specific guideline for the sense ordering in Wiktionary, we observed that the first entry is often the most frequently used one") which could be interesting and useful from a community perspective, but offers little actionable evidence and should be investigated further. The paper's conclusions identify some of the features that enable Wiktionary to rival expert-built lexicons: "We believe that its unique structure and collaboratively constructed contents are particularly useful for a wide range of dictionary users", listing eight such groups – among them "Laypeople who want to quickly look up the definition of an unknown term or search for a forum to ask a question on a certain usage or meaning."

On a critical note, the last paragraph says "we believe that collaborative lexicography will not replace traditional lexicographic theories, but will provide a different viewpoint that can improve and contribute to the lexicography of the future. Thus, Wiktionary is a rival to expert-built lexicons – no more, no less", which sounds a bit contradictory. The authors also note that "Lepore (2006: 87) raised a criticism about the large-scale import of lexicon entries from copyright-expired dictionaries such as Webster's New International Dictionary". It would be nice if the authors would write at least a short explanation of the problem that Lepore described. But the actual article[3] mentions Wiktionary only very briefly. For the most part, the article is a good academic-grade presentation of Wiktionary: it is very general and does not dive too much into details; it makes a few vague statements, but they present a good starting point for further research.

Wikipedia as an academic publisher?

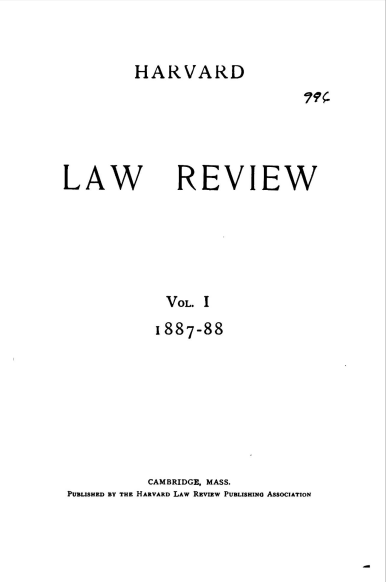
Xiao and Askin (2012) looked at whether academic papers could be published on Wikipedia.[4] The paper compares the publishing process on Wikipedia to that of an open-access journal, concluding that Wikipedia's model of publishing research seems superior, particularly in terms of publicity, cost and timeliness.

Can Wikipedia integrate with open-access scholarly publishing?

The biggest challenges for academic contributions to Wikipedia, they found, revolve around the level of acceptance of Wikipedia in academia, poor integration with academic databases, and technical and conceptual differences between an academic article and an encyclopedic one. However, the paper suffers from several problems. It correctly observes that the closest a Wikipedia article comes to a "final", fully peer-reviewed status is after having passed the featured article candidate process, but makes no mention of intermediary steps in Wikipedia's assessment project, such as B-class, Good Article and A-class reviews; nor is the the assessment project itself mentioned. Despite its focus on the featured-article process, no previous academic work on featured articles is cited (although quite a few have been published [5]). Crucially, the paper disregards the most relevant of Wikipedia's policies, no original research. Thus, the study fails to consider whether Wikipedia would want to publish academic articles without their undergoing changes to bring them closer to encyclopedic style – a topic that already has become an issue numerous times on the site, in particular regarding difficulties encountered by some educational projects. In the end, the paper, while a well-intentioned piece, seems to illustrate that university researchers can have a quite different understanding of what Wikipedia is than those more closely connected with the project.

In other news, however, a scientific journal appears to have found a viable way to publish peer-reviewed articles on Wikipedia: The open access journal *PLoS Computational Biology* has announced[6] that it is starting to publish "Topic Pages" - peer-reviewed texts about specific topics, which are published both in the journal and as a new article on Wikipedia. It is hoped that the Wikipedia versions will be updated and improved by the Wikipedia community. The first example is about circular permutation in proteins.

Wikipedia citations in American law reviews


HARVARD

LAW REVIEW

VOL. I
1887-88

CAMBRIDGE, MASS.
PUBLISHED BY THE HARVARD LAW REVIEW PUBLISHING ASSOCIATION

Volume 1 of the Harvard Law Review (1887–1888).

The article "A Jester's Promenade: Citations to Wikipedia in Law Reviews , 2002–2008" concerns the issue of citations of Wikipedia in US law reviews and the appropriateness of this practice.[7] The article seems to be well researched, and its author, law reference/research librarian Daniel J. Baker, demonstrates familiarity with the mechanics of Wikipedia (such as the permanent links). For the period 2002–08, Baker identified 1540 law-review articles that contain at least one citation of Wikipedia – most in law reviews dealing with general and "popular" subject matter, with a significant proportion originating from authors with academic credentials.

The article notes that 2006 marked the peak of that trend, attributing it (thereby demonstrating some familiarity with Wikipedia's history) to a delayed reaction to the Seigenthaler incident and the Essjay Controversy. (Since the article's data analysis ends in 2008, the question of whether this trend has rebounded in recent years is left unanswered.)

The author is highly critical of Wikipedia's reliability, arguing that a source that "anyone can edit" – and where much of the information is not verified – should not be used in works that may influence legal decisions. Thus Baker calls for stricter rules in legal publishing, in particular that Wikipedia should not be cited. In a more surprising argument, the paper suggests that if information exists on Wikipedia, it should be treated as common knowledge, and thus does not require referencing (a recommendation that follows a 2009 one – Brett Deforest Maxfield, "Ethics, politics and securities law: how unethical people are using politics to undermine the integrity of our courts and financial markets", 35 *OHIO N.U. L. REV.* 243, 293 (2009)). This argument does, however, raise the question of whether no citation at all is truly better than a citation to Wikipedia; if such a recommendation were followed, it could lead to a proliferation of uncited claims in law review journals that would be assumed (without any verification) to rely on "common knowledge" as represented in the "do not cite" Wikipedia.

One in four of articles tagged as flawed, most often for verifiability issues

A paper titled "A Breakdown of Quality Flaws in Wikipedia"[8] examines cleanup tags on the English Wikipedia (using a January 2011 dump), finding that 27.53% of articles are tagged with at least one of altogether 388 different cleanup templates. In a 2011 conference poster [9] (a version of which was summarized in an earlier edition of this newsletter), the authors analyzed – together with a third collaborator – a 2010 dump of the English Wikipedia for a smaller set of tags, arriving at much lower ratio: "8.52% [of articles] have been tagged to contain at least one of the 70 flaws". Using a classification of Wikipedia articles into 24 overlapping topic areas (derived from Category:Main topic classifications), the highest ratio of tagged articles were found in the "Computers" (48.51%), "Belief" (46.33%) and "Business" (39.99%) topics; the lowest were in "Geography" (19.83%), "Agriculture" (22.57%) and "Nature" (23.93%). Of the 388 tags on the more complete list, "307 refer to an article as a whole and 81 to a particular text fragment". As another original contribution of the paper, the authors offer an organization of the existing cleanup tags into "12 general flaw types" – the most frequent being "Verifiability" (19.46% of articles have been tagged with one of the corresponding templates), "Wiki tech" (e.g. the "orphan", "wikify" or "uncategorized" templates; 5.47% of articles) and "General cleanup" (2.01%).

Time evolution of Wikipedia discussions

Kaltenbrunner and Laniado look at the time evolution of Wikipedia discussions, and how it correlates to editing activity, based on 9.4 million comments from the March 12, 2010 dump.[10] Peaks in commenting and peaks in editing often co-occur (for sufficiently large peaks of 20 comments, 63% of the time) within two days. They show the articles with the longest comment peaks and most edit peaks, and the 20 slowest and 20 fastest discussions.

The authors note that a single, heavy editor can be responsible for edit peaks but not comment peaks; peaks in the discussion activity seem to indicate more widespread interest by multiple people. They find that "the fastest growing discussions are more likely to have long lasting edit peaks" and that some editing peaks are associated with event anniversaries. They use the Barack Obama article as a case study, showing peaks in comments and editing due to news events as well as to internal Wikipedia events (such as an editor poll or article protection). Current events are often edited and discussed in nearly real-time in contrast to articles about historical or scientific facts.

They use the h-index to assess the complexity of a discussion, and they chart the growth rate of the discussions. For instance, they find that the discussion pages of the three most recent US Presidents show a constant growth in complexity but that the rate of growth varies: Bill Clinton's talk page took 332 days to increase h-index by one, while George W. Bush's took only 71 days.

They envision more sophisticated algorithms showing the relative growth in edits and discussions. Their ideas for future work are intriguing – for instance, the question of how to determine article maturity and the level of consensus, based on the network dynamics. (AcaWiki summary [11])

APWeb2012 papers on admin networks, mitigating language bias and finding "minority information"

Several of the accepted papers [12] of this month's Asia-Pacific Web Conference APWeb2012 concerned Wikipedia:

- **Prototype tool searches for expert admins**: In the article "Exploration and Visualization of Administrator Network in Wikipedia",[13] four Chinese authors examine the collaboration graph of administrators on the English Wikipedia (where two of them are connected by an edge if they have edited the same article during the sampled time span from January 2010 to January 2011), and "define six features to reflect the characteristics of administrator's work from different respects including diversity of the admin user interests, the influence & importance across the network, and longevity & activity in terms of contribution." The authors observe that the recognition of an admin's work by other users in the form of barnstars seems to agree with the overall rank they calculate from these quantities: "By analyzing the profiles of the top ranked fifty admin users as a test case, it has been observed that the number of barn stars received by them also follows the similar trend as we overall ranked the admin users."To extract topics from an admin's history and define diversity, Latent Dirichlet allocation (LDA) is used. The authors describe a prototype software called "Administrator Exploration Prototype System", which displays these various quantitative measures for an admin and allows ranking them. In particular, it "will automatically find the expert authors based on the editing history of each admin user". An example screenshot shows a list of results for a "Search for "Expert Admin User" for the keywords "Music, Songs, Singers", topped by Michig, Mike Selinker and Bearcat. Analyzing the whole network, the authors find a "decreasing trend of the clustering coefficient [which] can also be seen as a symptom of the growing centralization of the network." Overall, they observe that "the administrator network is a healthy small world community having a small average distances and a strong centralization of the network around some hubs/stars is observed. This shows a considerable nucleus of very active administrators who seems to be omnipresent."

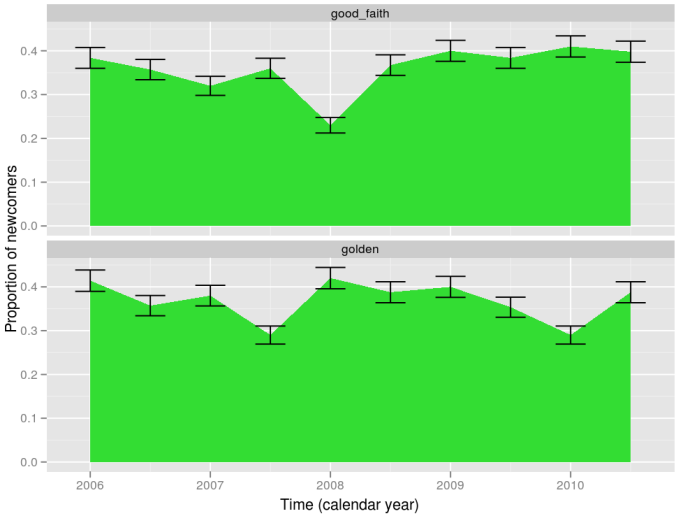
Detecting "minority information" on Wikipedia: A paper[14] by two Japanese researchers proposes "a method of searching for minority information that is less-acknowledged and has less popularity in Wikipedia" for a given keyword. "For example, if the user inputs 'football' as a majority information keyword, then the system seeks articles having a sentence of "....looks like football...." or similar content of articles about soccer in Wikipedia. It extracts as candidates for minority sports those articles which have few edits and few editors. Then, it performs sports filtering and extracts minority articles from the candidates. In this case, the results are 'Bandy', 'Goalball', and 'Cuju'." The authors constructed a prototype system and tested it.

Not what the majority of readers search for under "football": A goalball game

- **Completing Wikipedia articles with information from other language versions**: In an article titled "Extracting Difference Information from Multilingual Wikipedia"[15], four Japanese researchers describe a "method for extracting information which exists in one language version [of Wikipedia], but which does not exist in another language version. Our method specifically examines the link graph of Wikipedia and structure of an article of Wikipedia. Then we extract comparison target articles of Wikipedia using our proposed degree of relevance." As motivating example, they note that the English Wikipedia's coverage of the game cricket is much fuller than the Japanese Wikipedia's, but spread over separate articles beyond just the main one at cricket. The goal is a system where a (Japanese) user can enter a keyword and will receive the "Japanese article with sections of English articles that do not appear in the Japanese article".

Briefly

Unchanged quality of new user contributions over time. GroupLens PhD candidate Aaron Halfaker (who also collaborates with the Wikimedia Foundation as a contractor research analyst) shared some preliminary results on the quality of new user contributions,[16] part of a larger study currently submitted for publication. The results, based on an analysis of revert rates in the English Wikipedia combined with blind assessment of a new editor contribution history, indicate that new editors have produced the same level of quality in their first contributions since 2006. Despite the fact that "the majority of new editors are not out to obviously harm the encyclopedia (~80 percent), and many of them are leaving valuable contributions to the project in their first editing session (~40 percent)", today's user experience for a first-time editor is much more hostile than it used to be, as "the rate of rejection of all good-faith new editors' first contributions has been rising steadily, and, accordingly, retention rates have fallen. These results challenge the hypothesis that today's newbies produce much lower quality contributions than in earlier years.



The proportion of "good faith" and "golden" editors among new contributors over time has remained constant.[16]

- **Modeling Wikipedia's community formation processes**. An important factor behind the success of Wikipedia is its own internal culture. Like any social group, a community of peer production has its own . Unlike traditional social groups -- a recently-defended doctoral dissertation in computer science argues -- the process of formation

of these traits involves, and often determines, how contents are being produced. The dissertation, defended by former Summer of Research fellow and Wikimedia Foundation contractor analyst Giovanni Luca Ciampaglia,[17] uses computer simulation to study how the community of Wikipedia may have formed its specific cultural traits and distinctive sociological features. Starting from the distribution of user account lifespan in five of the largest Wikipedia communities (English, German, Italian, French, and Portuguese) this work shows how the statistical patterns of the data can be reproduced by a simple model of cultural formation based on principles taken from self-categorization theory and social judgment theory.

The research finds that an important factor to determine whether a community will be able to sustain itself and thrive is the degree of openness of individual users towards differing points of view, which may be critical in the early stages of user participation, when a newcomer first enters in contact with the body of social norms that the community has devised. The thesis concludes that simulation techniques, when supplemented with empirical methods and quantitative calibration, may become an important tool for conducting sociological studies.

Distribution of AFT ratings for articles in different project quality assessment categories.[18]

- **Matching reader feedback via the Article Feedback Tool to editor peer review**: An upcoming presentation at Wikimania 2012[18] compares data gathered from the Article Feedback Tool (AFT) version 4 on the English Wikipedia over summer 2011 to ratings assigned by various peer review processes, e.g. good and featured articles. As might be expected, articles at any point in the peer review process tend to be rated more highly by reviewers, but this distinction is highly sensitive to the article length. Once length is accounted for (using a variety of methods), the differences between demoted or not promoted articles and unrated articles disappears. The research also offers a broad snapshot of the AFT dataset as well as some suggestions for future AFT design. Future revisions of the draft as well as the presentation will approach the dynamic relationship between peer reviewed status and reader feedback, exploiting entry and exit into various categories for identification.
- **Referencing of Wikipedia in academic works is continuing unabated**: An article in the "Research Trends" newsletter published by the bibliographical database Scopus, titled "The influence of free encyclopedias on science"[19] charts the number of papers in Scopus that are either about Wikipedia or cite it. Considering that Wikipedia was only founded in 2001 (i.e. that these numbers have necessarily started from zero right before the observed timespan), the author's astonishment at the compound annual growth rates for both kinds of papers from 2002 to 2011 (which she calls "staggering" and "unbelievable", respectively) is somewhat surprising, but the article also gives the growth rates for the five years from 2007 to 2011 (ca. 19% per year for Wikipedia as a subject, ca. 31% per year for Wikipedia as a reference). Interestingly, Scholarpedia is showing itself to be the second most popular online encyclopedia to be cited, if lagging significantly behind Wikipedia (5%).
- **Using Wikipedia to drive traffic to library collections**: In an article titled "Wikipedia Lover, Not a Hater: Harnessing Wikipedia to Increase the Discoverability of Library Resources"[20] in the *Journal of Web Librarianship*, two librarians from the University of Houston Libraries and a former intern report how they had successfully used Wikipedia to drive traffic to the collection of the institution's digital services department (UHDS), proceeding from merely inserting links into articles to uploading images from the collection to Commons (which still contain such link on the file description pages): "Originally, UHDS intended to contribute

exclusively to the External Links section of existing Wikipedia articles. [However, over time] UHDS staff found it was much more effective to match digital items with Wikipedia articles and to share those items in Wikimedia Commons (WMC) rather than (or in addition to) the External Links section of the articles." While few statistics are given, the authors emphasize the effectiveness of their actions, observed already for the very first attempts: "Within hours of posting external links to existing Wikipedia articles, the digital library received hits to those collections at a surprisingly high rate." As an example of an article enriched with such images, the entry 1915 Galveston hurricane is named. Among the successful additions to external links section is the article about former US president George H. W. Bush, where the student intern linked [21] a photograph [22] showing Bush shaking hands with former University of Houston chancellor Philip G. Hoffman (as already noted in the Signpost's April 2011 coverage after the authors had presented their project at the annual meeting of the Association of College and Research Libraries: "Experts and GLAMs – contributing content or 'just' links to Wikipedia?"). Much of the paper describes basic technicalities of Wikipedia: The uploading of image, the use of contributions lists, talk pages and watchlists. While Wikipedia's external links guidelines are not cited in the paper, it notes that "contributing effectively to Wikipedia and WMC entailed a steep learning curve in order to align contributions with the granular and well-enforced Wikipedia guidelines for use", and among them notices policies against advertising. As one unresolved problem for such institutional usage of Wikipedia and Commons, the paper describes the prohibition "to share an editor username with other editors, and [that] organizational usernames are considered a violation of Wikipedia guidelines forbidding the promotion of organizations. When the pilot project transitioned into a permanent departmental program, UHDS staff struggled to devise a way that others on staff could continue to monitor previous edits and uploads and create new ones", e.g. due to the lack of shared watchlists.

- **Weekly and daily activity patterns discern Wikipedia from commercial sites**: Two Finnish researchers analyzed[23] the distribution of timestamps in the recent changes RSS feed from four different language versions (Arabic, Finnish, Korean, and Swedish - Arabic having been chosen because its speakers are spread over "a very wide range of timezones", in contrast to the other three), and RSS new feeds from BBC World News "and the leading Finnish daily newspaper Helsingin Sanomat". As the main difference between the activity on those two sites (which the authors describe as "commercial news sites") and on the Wikipedias, it was found that Wikipedia edits "distribute fairly equally over all days in all cases. The drop of activity on weekends that occurred with the commercial news services is not visible in the Wikipedias, quite the opposite, with Sundays typically seeing the highest average level of activity. Only the Arabic version has a slightly lower activity rate in Sundays, however, we should remember the fact that in Arabic countries the weekend falls on Friday-Saturday or in some countries on Thursday-Friday". The diurnal patterns are found to be "more spread out" on Wikipedia, where "the activity levels follow natural diurnal rhythms. Interestingly, a great number of changes are made during working hours, which leads us to 2 different, but not mutually exclusive, conjectures about the people who edit Wikipedia. Either, the editors are people with “free” time during the day, e.g., students, or people actually edit Wikipedia during the working hours at work. Our methodology is not able to answer this question".
  Furthermore, the authors offer a rather far-reaching but (if proven) significant conjecture based on their date: "Cultural and geographical differences in the Wikipedias we studied seemed to have very little effect on the level of activity. This leads us to speculate that the 'trait' of editing Wikipedia is something to which individuals are drawn, not something specific to certain cultures."
  Last year, papers by two other teams (covered in the September issue of this newsletter: Wikipedians' weekends in international comparison", but missing from the "Related work" section of the present paper) had similarly examined daily and weekly patterns on Wikipedia, coming to other results - in particular, different language Wikipedias showed different weekly patterns.
- **Simple English Wikipedia is only partially simpler/controversy reduces complexity**: "A practical approach to language complexity: a Wikipedia case study"[24] analyzed samples of articles from the English Wikipedia and the Simple English Wikipedia from the end of 2010 with respect to the Gunning fog index as well as other

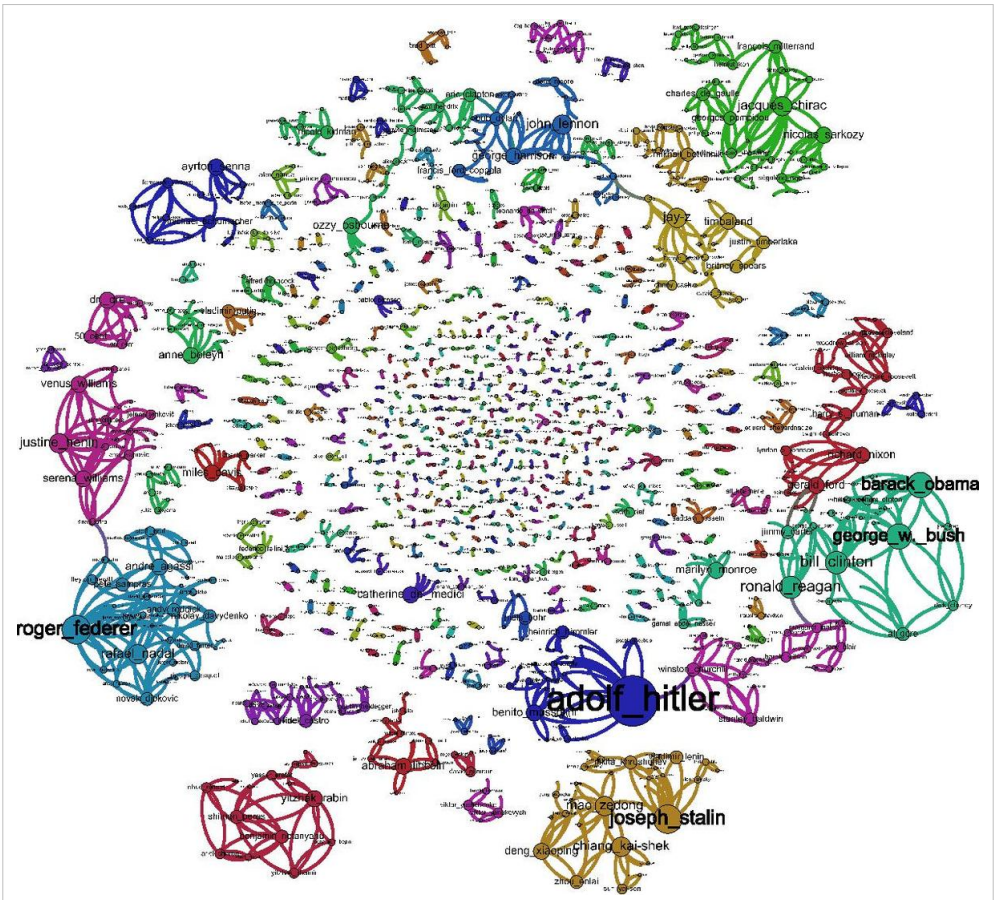
measures for language complexity. Comparing them with other corpora including Charles Dickens' books, they observe that "Remarkably, the fog index of Simple English Wikipedia is higher than that of Dickens, whose writing style is sophisticated but doesn't rely on the use of longer latinate words which are hard to avoid in an encyclopedia. The British National Corpus, which is a reasonable approximation to what we would want to think of as 'English in general' is a third of the way between Simple and Main, demonstrating the accomplishments of Simple editors, who pushed Simple half as much below average complexity as the encyclopedia genre pushes Main above it." However, the number of distinct tokens used (a measure for vocabulary richness) is almost the same on the English and Simple Wikipedia (the samples were chosen to be of the same size). Still "detailed analysis of longer units (n-grams rather than words alone) shows that the language of Simple is indeed less complex". In another finding, the authors "investigate the relation between conflict and language complexity by analysing the content of the talk pages associated to controversial and peacefully developing articles, concluding that controversy has the effect of reducing language complexity."

- **Contributions from South America**. "Mapping Wikipedia edits from South America",[25] the latest from a series of studies and visualizations by Oxford Internet Institute researcher Mark Graham and his team, reports that almost half of all edits to Wikipedia from South America come from Brazil, which is unsurprising considering that the largest population of Internet users in South America lives in Brazil. More interestingly, Chile ─- a country with only 5-6% of the continent's Internet population -- contributes more than 12% of edits to Wikipedia.
- **Deaths generate edit bursts**: A student paper titled "Death and Change Tracking : Wikipedia Edit Bursts"[26] examines the editing activity in nine articles about celebrity actors on the English Wikipedia after they died.
- **Searching by example**: This month's WWW 2012 conference in Lyon, France saw a demo titled "SWiPE: Searching Wikipedia by Example"[27], showcasing a tool where the user can search for articles similar to a given one by modifying entries in that article's infobox, and also ask questions in natural language.
- **Wikipedia in the eyes of PR professionals**. A study published in the journal of the Public Relations Society of America (PRSA)[28] surveyed public relations and communications professionals about their perception of Wikipedia contribution and conflict of interest. The online survey was pilot-tested with members of the Corporate Representatives for Ethical Wikipedia Engagement (many of whom have recently pushed for Wikipedia to let PR professionals edit articles about their clients to a greater extent) and produced 1,284 usable responses after being disseminated via various outlets. The results indicate that "of the 35% who had engaged with Wikipedia, most did so by making edits directly on the Wikipedia articles of their companies or clients". The response time to issues reported on talk pages was found to be one of the important barriers in the interaction between Wikipedia community members and PR professionals. The author observes that "when the wait becomes too long, the content is defamatory, or a dispute with a Wikipedian needs to be elevated, there are resources to help. Unfortunately, only a small percentage of the respondents in this study had used them and many had never heard of these resources". As another argument against the "bright line" rule advocated by Wikipedia's Jimmy Wales (which says that PR professionals should not edit Wikipedia articles they are involved in), a separate result of the paper has been offered, which has met with heavy criticism by Wikimedians regarding statistical biases and other issues (see e.g. last week's Signpost coverage: "Spin doctors spin Jimmy's 'bright line'"): 32% of the respondents said that "there are currently factual errors on their company or client's Wikipedia articles", corresponding to 41% of those respondents who said that such articles existed, or 60% of those respondents who said that such articles existed but did not reply "don't know" to that question. The press releases of the author's college [29] and of PRSA [30] interpreted the result as "Sixty percent of Wikipedia articles about companies contain factual errors", although the latter was updated [31] after the criticism "to clarify the survey findings described in this press release and help prevent any misinterpretation of the data that this release may have caused".
- **Wikipedia coverage of marketing terms found accurate**: The proceedings of the recent "International Collegiate Conference Faculty" of the American Marketing Association (AMA) offer a more positive view on Wikipedia from PR professionals: "Is Wikipedia A Reliable Tool for Marketing Educators and Students? A Surprising Heck Yes!".[32] The paper chose a more systematic way to examine the quality of Wikipedia articles

than the PRSA study and focused on AMA's area of expertise, starting out from a "random sample of marketing glossary terms [that] were collected from 3 marketing management textbooks and 4 marketing principles textbooks", and rating corresponding Wikipedia entries from 1 to 3 according to a standard procedure for content analysis: "Each textbook definition was compared to the corresponding Wikipedia definition and rated using a 3-point Likert scale where 1=Correct Definition, 2=Correct but difficult to find the term or the definition was not easy to decipher, or 3=Incorrect definition when compared to the textbook term.". Of 459 items in the eventual sample only five were rated 3, and "the average score across all textbooks was a 1.18 demonstrating Wikipedia is an accurate source of marketing content."

- **Wikipedia's osteosarcoma coverage assessed**: An abstract published in the Journal of Bone & Joint Surgery[33] finds "that the quality of osteosarcoma-related information found in English Wikipedia is good but inferior to the patient information provided by the National Cancer Institute". The abstract refers to a study and results that appear to be identical to the one reported in a 2010 viewpoint article in the Journal of the American Medical Informatics Association (JAMIA) (*Signpost* coverage).
- **Wikipedia assignments for Finnish school students**: A paper by three Finnish authors[34] describes course assignments to upper secondary school students (age 16–18) involving "writing articles for Wikipedia (a public wiki) and for the school's own wiki", in subject areas including biology, geography and Finnish history. In particular the paper reports that "a carefully planned library [visit] can help to activate students to use printed materials in their source-based writing assignments. [And that our] findings corroborate the generally held view that students tend to copy-paste and plagiarise, especially when exploiting Web sources."
- **Wikipedia as a thermodynamic system - becoming more efficient over time**: A paper titled "Thermodynamic Principles in Social Collaborations"[35] (presented at this month's Collective Intelligence 2012 [36] conference) applies principles and concepts from Statistical mechanics to the collaboration on (the English) Wikipedia. The analogy is based on interpreting the edit count of a user as the energy level of a particle, positing a "logarithmic energy model" for edits which assumes a "decreasing effort required for a given user to make additional edits in a relatively short period of time (e.g., one month) or to a particular page". (According to the authors this contrasts with two other theories which also explain the observed power law distribution of edit counts: The "Wikipedia editors are 'born'" notion, which assumes that different users need to expend different amounts of energy on the same kind of edits due to "an extreme heterogeneity of preference among the potential user population", and the "Wikipedia editors are 'made'" notion, which sees positive or negative feedback from other users as the defining influence.) Using the analogy, the authors define the entropy, free energy, temperature, entropy efficiency and entropy reduction of an editing community and their edits during a particular timespan. They then calculate the latter two for each month in the English Wikipedia's history from January 2002 to December 2009. They conclude that "Wikipedia has become more efficient in terms of entropy efficiency, and more ordered according to entropy reduction. The increasing power-law coefficient causes the shift of the contributions from elites to crowd. The saturation of free energy reduction ratio may cause the saturation of the active editors." The next section finds that "entropy efficiency is correlated with the quality of the social collaboration", and one figure is interpreted as implying "that the nature of Wikipedia is a true media of the masses, where pages produced by crowd wisdom will have higher quality and thus more readership compared to that produced by a few elites."
- **Too many docs don't spoil the broth**: Another paper[37] presented at the Collective Intelligence 2012 conference similarly found "that the number of contributors has a curvilinear relationship to information quality, more contributors improving quality but only up to a certain point" - based on an examination of 16,068 articles in the realm of the WikiProject Medicine.
- **The most influential biographies vary depending on the language/culture**: Barcelona Media Foundation studied "the most influential characters" in the 15 largest language Wikipedias[38], by asking which biographies are the most linked to ("central") from other Wikipedia biography articles. Political and artistic biographies are the most central, and the particular biographies depend on the language. They found, for instance, that Shakespeare's biography is among the most important for Russian, Chinese, Spanish, and Dutch, but not for

English. And they estimated the Jaccard similarity coefficient (similarity) between the social networks in different language editions: most similarity can be explained by language-family and geographical or historical ties. One interesting finding is that Dutch "seems to serve as a bridge between different language and cultural groups". Some social connections are very common,and they produce a graph of the connections found in at least 13 of the language editions [39]. The authors note that articles on people from non-Anglo-Saxon cultures may be missing if they are not known internationally, since the initial list of notable people is extracted from DBPedia. A blog post on *Technology Review* highlighted the fact that in the paper's table of most connected biographies (listing the top 5 from 15 language versions), among the 75 entries "only three are women: Queen Elizabeth II, Marilyn Monroe and Margaret Thatcher" , which it interprets as one of "The Worrying Consequences of the Wikipedia Gender Gap [40]". (Summary at AcaWiki [41])



Biographical social network of the connections between persons present in at least 13 of the 15 largest language Wikipedias, as described in Aragón et al.[38]

References

[1] Restivo, M. & van de Rijt, A. (2012). Experimental Study of Informal Rewards in Peer Production. *PLoS ONE* 7(3): e34358. **PDF** (http://www.plosone.org/article/fetchObjectAttachment.action?uri=info:doi/10.1371/journal.pone.0034358&representation=PDF) • **DOI** Open access

[2] Meyer, C. M., & Gurevych, I. (2012). Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography*. Oxford: Oxford University Press. **PDF** (https://www.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2011/oup-elex2012-meyer-wiktionary.pdf) Open access

[3] Lepore, J. (2006). Noah's Mark (http://www.newyorker.com/archive/2006/11/06/061106fa_fact_lepore), *The New Yorker*, November 6, 2006, pp. 78-86. **HTML** (http://www.readliterature.com/noahwebster.htm) Open access

[4] Xiao, L., & Askin, N. (2012). Wikipedia for Academic Publishing: Advantages and Challenges. *Online Information Review*, 36(3), 2. Emerald Group Publishing Limited. **HTML** (http://www.emeraldinsight.com/journals.htm?articleid=17026842&show=abstract) Closed access

[5] http://wikilit.referata.com/wiki/Category:Featured_articles

[6] (2012) "Topic Pages: PLoS Computational Biology Meets Wikipedia". *PLoS Computational Biology* 8 (3): e1002446. : 10.1371/journal.pcbi.1002446 (http://dx.doi.org/10.1371/journal.pcbi.1002446). Open access

[7] Baker, D. J. (2012). A Jester's Promenade: Citations to Wikipedia in Law Reviews, 2002–2008. *I/S: A Journal of Law and Policy for the Information Society*, 7(2):1–44. **PDF** (http://moritzlaw.osu.edu/students/groups/is/files/2012/02/Baker.FE_.Final_.Weber_.pdf) Open access

[8] Anderka, M., & Stein, B. (2012). A breakdown of quality flaws in Wikipedia. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality – WebQuality '12* (p. 11). New York: ACM Press. **DOI** (http://dx.doi.org/10.1145/2184305.2184309) • **PDF** (http://www.uni-weimar.de/medien/webis/publications/papers/stein_2012d.pdf) Open access

[9] Anderka, M., Stein, B., & Lipka, N. (2011). Towards automatic quality assurance in Wikipedia. *Proceedings of the 20th international conference companion on World Wide Web – WWW '11*. New York: ACM Press. **DOI** (http://dx.doi.org/10.1145/1963192.1963196) • **PDF** (http://www2011india.com/proceeding/companion/p5.pdf) Open access

[10] Kaltenbrunner, A., & Laniado, D. (2012). There is No Deadline – Time Evolution of Wikipedia Discussions. *ArXiV*. Computers and Society; Physics and Society. **PDF** (http://arxiv.org/abs/1204.3453) Open access

[11] http://acawiki.org/There_is_no_deadline_-_Time_evolution_of_Wikipedia_discussions

[12] http://e-research.csm.vu.edu.au/files/apweb2012/acceptedpapers.html

[13] Yousaf, J., Li, J., Zhang, H., & Hou, L. (2012). Exploration and Visualization of Administrator Network in Wikipedia. In: Q. Z. Sheng, G. Wang, C. S. Jensen, & G. Xu (Eds.), *Web Technologies and Applications*, Lecture Notes in Computer Science 7235:46-59. Berlin, Heidelberg: Springer Berlin Heidelberg. **DOI** (http://dx.doi.org/10.1007/978-3-642-29253-8_5) Closed access

[14] Hattori, Y., & Nadamoto, A. (2012). Search for Minority Information from Wikipedia Based on Similarity of Majority Information. In: Q. Z. Sheng, G. Wang, C. S. Jensen, & G. Xu (Eds.) *Web Technologies and Applications*, Lecture Notes in Computer Science 7235:158-169. Berlin, Heidelberg: Springer Berlin Heidelberg. **DOI** (http://dx.doi.org/10.1007/978-3-642-29253-8_14) Closed access

[15] Fujiwara, Y., Suzuki, Y., Konishi, Y., Nadamoto, A., Sheng, Q., Wang, G., Jensen, C., et al. (2012). Extracting Difference Information from Multilingual Wikipedia. In: Q. Z. Sheng, G. Wang, C. S. Jensen, & G. Xu (Eds.) *Web Technologies and Applications*, Lecture Notes in Computer Science 7235:496-503. Berlin, Heidelberg: Springer Berlin Heidelberg. **DOI** (http://dx.doi.org/10.1007/978-3-642-29253-8_42)Closed access

[16] Halfaker, A. (2012). Kids these days: the quality of new Wikipedia editors over time. *Wikimedia Foundation blog*. **HTML** (http://blog.wikimedia.org/2012/03/27/analysis-of-the-quality-of-newcomers-in-wikipedia-over-time/) Open access

[17] Ciampaglia, G. L. (2011). *User participation and community formation in peer production systems*. PhD Thesis, Università della Svizzera Italiana **PDF** (http://doc.rero.ch/record/28987) Open access

[18] Hyland, A. (2012). Comparing article quality by article class and article feedback ratings. *Wikipedia*. **HTML** (http://en.wikipedia.org/wiki/User:Protonk/Article_Feedback) Open access

[19] Huggett, S. (2012). The influence of free encyclopedias on science. *Research Trends*, (27). **HTML** (http://www.researchtrends.com/issue-27-march-2012/the-influence-of-free-encyclopedias-on-science/)Open access

[20] Elder, D., Westbrook, R. N., & Reilly, M. (2012). Wikipedia Lover, Not a Hater: Harnessing Wikipedia to Increase the Discoverability of Library Resources. *Journal of Web Librarianship*, 6(1), 32-44. Routledge. **DOI** (http://dx.doi.org/10.1080/19322909.2012.641808) Closed access

[21] http://en.wikipedia.org/w/index.php?title=George_H._W._Bush&diff=399031384&oldid=397526689

[22] http://digital.lib.uh.edu/cdm4/document.php?CISOROOT=/p15195coll6&CISOPTR=226&CISOSHOW=224

[23] Karkulahti, O., & Kangasharju, J. (2012). Surveying Wikipedia activity: Collaboration, commercialism, and culture. *The International Conference on Information Network 2012* (pp. 384-389). IEEE. **DOI** (http://dx.doi.org/10.1109/ICOIN.2012.6164405) Closed access

[24] Yasseri, T., Kornai, A., Kertész, J. (2012). A practical approach to language complexity: a Wikipedia case study. *ArXiv*. Computation and Language. **PDF** (http://arxiv.org/abs/1204.2765) Open access

[25] Graham, M. (2012). Mapping Wikipedia edits from South America. *Zero Geography*. **HTML** (http://www.zerogeography.net/2012/04/mapping-wikipedia-edits-from-south.html) Open access

[26] Lincoln, M. (2012). *Death and Change Tracking : Wikipedia Edit Bursts*. **PDF** (http://people.lis.illinois.edu/~mlincol2/content/deathedits.pdf) Open access

[27] Atzori, M., & Zaniolo, C. (2012). SWiPE: Searching wikipedia by example. *Proceedings of the 21st international conference companion on World Wide Web* - WWW '12 Companion (p. 309). New York, New York, USA: ACM Press. **DOI** (http://dx.doi.org/10.1145/2187980.2188036) • **PDF** (http://swipe.i-mozart.com/www2012demo/TR_swipe.pdf) Open access

[28] DiStaso, M. W. (2012). Measuring Public Relations Wikipedia Engagement: How Bright is the Rule? *Public Relations Journal*, 6(2) **HTML** (http://www.prsa.org/Intelligence/PRJournal/Spring_12/) Open access

[29] http://www.eurekalert.org/pub_releases/2012-04/djc-sfm041712.php

[30] http://media.prsa.org/article_display.cfm?article_id=2575

[31] http://media.prsa.org/article_display.cfm?article_id=2582

[32] Gray, D. M., & Peltier, J. (2012). Is Wikipedia Reliable Tool for Marketing Educators and Students? A Surprising Heck Yes! *Marketing Always Evolving. 34th Annual International Collegiate Conference*. **PDF** (http://www.marketingpower.com/Community/collegiate/Documents/FY12+Documents/2012+Faculty+Proceedings.pdf) Open access

[33] Leithner, A., Maurer-Ertl, W., Glehr, M., Friesenbichler, J., Leithner, K., & Windhager, R. (2012). Wikipedia and Osteosarcoma: An educational opportunity and professional responsibility for Emsos. *Journal of Bone & Joint Surgery*, BR, 94-B(SUPP XIV), 13. British

Editorial Society of Bone and Joint Surgery. **HTML** (http://www.bjjprocs.boneandjoint.org.uk/content/94-B/SUPP_XIV/13.abstract) Closed access

[34] Sormunen, E., Eriksson, H., & KurkipaÌa, T. (2012). Wikipedia and wikis as forums of information literacy instruction in schools. *The Road to Information Literacy: Librarians as Facilitators of Learning*. IFLA 2012 Congress Satellite Meeting (pp. 1-23). **PDF** (https://www12.uta.fi/blogs/know-id/files/2012/04/Sormunen_Eriksson_TuulipÃ¤Ã¤_final.pdf) Open access

[35] Peng, H.-K., Zhang, Y., Pirolli, P., & Hogg, T. (2012). Thermodynamic Principles in Social Collaborations. *ArXiV*. Physics and Society. **PDF** (http://arxiv.org/abs/1204.3663) Open access

[36] http://www.ci2012.org/accepted-papers

[37] Kane, G. C., & Ransbotham, S. (2012). Collaborative Development in Wikipedia. *ArXiv*. **PDF** (http://arxiv.org/abs/1204.3352) Open access

[38] Aragón, P., Kaltenbrunner, A., Laniado, D., & Volkovich, Y. (2012). Biographical Social Networks on Wikipedia - A cross-cultural study of links that made history. *ArXiV*. Computers and Society; Physics and Society, **PDF** (http://arxiv.org/abs/1204.3799) Open access

[39] http://produccionmultimedia.barcelonamedia.org/var/www/public_html/wp-content/uploads/2012/04/intersection13_boundedV2.pdf

[40] http://m.technologyreview.com/blog/arxiv/27777/

[41] http://acawiki.org/Biographical_Social_Networks_on_Wikipedia_-_A_cross-cultural_study_of_links_that_made_history

Issue 2(5): May 2012

Supporting interlanguage collaboration; detecting reverts; Wikipedia's discourse, semantic and leadership networks, and Google's Knowledge Graph

With contributions by: Jodi.a.schneider, Piotrus, Tbayer and Angelika Adam

Discourse on Wikipedia sometimes irrational and manipulative, but still emancipating, democratic and productive

An article[1] in sociology journal *The Information Society* looks at interactions between Wikipedia editors and the project's governance, visible in the articles on stem cells and transhumanism, and in the analysis of Wikipedia's discussion of userboxes, all through the prism of Jürgen Habermas universal pragmatics and Mikhail Bakhtin dialogism theories.

The authors focus on the qualitative analysis of language used by editors, to argue that Wikipedia has elements of a democracy, and is an example of a Web 2.0–empowering discourse tool. They stress that some forms of discourse found online (including on Wikipedia) may be highly irrational, something that some previous arguments that Web 2.0 is a democratic space have often ignored, but they argue that this is in fact not as much of a hindrance as previously expected. Cimini and Burr remark that discourse can develop between Wikipedians of widely differing points of view, and that some editors will engage in "repeated, strategic, and often highly manipulative attempts" to assert personal authority. Such discussions may be very lively, involving "personal, emotional, or humour-based arguments", yet the authors argue that such comments may not be a hindrance; instead, "on many occasions, there is thus a clearer exposition of views that is achieved, in spite of, or perhaps because of, these personal [and] sometimes vulgar methods of argumentation."

In the end, the authors are positive about the success of Wikipedia's deliberation in reaching consensus, although they say that it can be "fleeting and transitory" on occasion. Unfortunately, the paper does not touch on Wikipedia policies such as Wikipedia:Civility and Wikipedia:No personal attacks, which would certainly have added to their analysis.

Despite the paper's claim to have received approval for research through a university research ethics committee, the paper does critically discuss the postings of specifically named editors ("[Editor A's] claim to authority and *ad hominem* attacks were met with derision by [Editor B]" (names replaced by the *Signpost*); this may raise eyebrows. Not all editors are 100% anonymous, which raises the question of whether the researchers did enough to protect the identity and reputation of the editors it cites. At the very least, why weren't the editors' usernames changed in the quotes? Their direct identification adds nothing to the article, and may expose the users to attack. (Similar questions

have been discussed in the past by members of the Wikimedia Foundation Research Committee.)

Different language Wikipedias: automatic detection of inconsistencies

In a paper presented at the *4th International Conference on Intercultural Collaboration* (ICIC),[2] Kulkarni et al. offer a simple approach to support the work of Wikipedia editors who maintain articles concerning the same topic in multiple language versions. The long-term goal is to implement a bot that supports these specialized users by highlighting missing attributes and content inconsistencies.

The analysis was focused on a pairwise comparison of infoboxes in different languages. First, the attribute-value pairs were extracted from the infoboxes and translated into English via Google translate. The identification of matching attribute names was achieved through direct text comparison with a set of synonyms obtained from *WordNet* (this step was included to handle mismatches caused by translation errors and variations). In a second step (the matching of attribute-values) the authors again used direct text comparative methods, and checked whether the values could be identified as homophones, to exclude mismatches caused by spelling mistakes in the text.

The evaluation data-set of these analyses and the whole pipeline included articles from English, German, Chinese and Hindi Wikipedias concerning two restricted domains: Indian cities and US-based companies. The evaluation revealed "a significant increase in recall after the concepts of homophones and synonyms were applied in addition to the direct text comparison." But the overall result was very weak, mainly due to translation errors. The authors noticed syntactic and semantic differences between the infoboxes, such as paraphrasing or different fact representations. "Also, abbreviations, units conversion and geographic location matching [was not handled by their system]." The researchers plan to improve the system by addressing all of these issues in turn.

Finding deeper meanings from the words used in Wikipedia articles

An undergraduate computer science honors thesis at Trinity University constructs a semantic graph from 451 articles, linked to from the World War II article.[3] Ryan Tanner's goal is to produce a visualization "which allows one to quickly find and examine connections between the people, places and things described in Wikipedia". The process is as follows:

1. Import SQL dump from the Wikimedia Foundation into a local database
2. Strip wiki markup from the articles using Bliki
3. Parse articles with the Stanford NLP, using dependency grammars to extract facts and simplify sentences
4. Parse the output from the Stanford library using Scala
    1. Read a Stanford XML file into a collection of models.
    2. Produce abstractions for named entities and locations.
    3. Input models into the algorithm developed for this thesis (see Chapter 7)
5. Store results in a database.
6. Traverse the resulting graph and produce user-presentable output.

Originally the goal was to visualize the whole of Wikipedia; however, due to problems with the dump, only 250,000 articles out of about 1.5 million were imported. An even smaller subset was ultimately usable, since the Stanford NLP library crashed on many of the remaining articles due to markup issues and the need for manual cleanup. To ensure a dense graph, tests were focused on the network of the World War II article. Some brief examples of the resulting graph are given in Chapter 10, which notes false positives as one problem requiring further investigation. The author makes suggestions for future research, such as using Simple English Wikipedia [4] or more complex relations.

How leaders emerge in the Wikipedia community

A paper titled "Leading the Collective: Social Capital and the Development of Leaders in Core-Periphery Organizations"[5] looks at how leaders emerge in Wikipedia and similar crowd-based organizations. While often seen as egalitarian and with little hierarchy, such projects always have a group of leaders who have emerged from the community (the "crowd"), involved in planning, mediation, and policy development. The authors treat Wikipedia and similar organization as a core–periphery network model developed by Steve Borgatti—a system with a deeply interconnected center and a poorly connected periphery. In Wikipedia, the leaders ("core") comprise the most active contributors, and the authors assume they produce the most social capital. Using social network analysis, the paper looks at the interpersonal ties between the editors, focusing on the ties between leaders and periphery. The hypothesis is that specific types of ties will have a greater influence on advancement to leadership.

The authors collected data from RfA pages, and the ties were measured through user-talk-page interactions. Leaders were defined as admins, and periphery editors as non-administrators; this operationalization may raise some doubts about the validity, since some very active and prominent members of the community are not admins, something the authors do not address. The authors find that the most important ties are the early ones to the periphery, and later, ties to the leaders. Overall strong ties are not as important as weak ties, although Simmelian ties (between pairs of leader groups) are among the most important.

Collier and Kraut conclude that leaders in projects such as Wikipedia do not suddenly appear; instead, they evolve over time through their immersion in the project's social network. Early in their experience, those leaders gain a deeper understanding of the community, developing a network of contacts through their weak ties to the periphery; later, their most important ties are to the leaders, particularly in the form of strong connection to a leader group.

Identifying software needs from Wikipedia translation discussions

A paper[6] presented at an international conference on intercultural collaboration aims "to identify the type of community interaction needed for successfully creating or amending an article via Wikipedia translation activities", and proposes new software tools to facilitate these interactions. To this end, the researchers from Kyoto University analyzed 1694 talk-page comments from three Wikipedias, belonging to articles in categories marking (partial or complete) translations (e.g. fr:Catégorie:Projet:Traduction/Articles_liés): 228 articles from the Finnish, 93 from the French, and 94 from the Japanese Wikipedia. They attempted to categorize (code) each comment according to which "activity" it referred to (either editing the article or translating it), about which "context" it was referring to (using the categories "content", "layout", "sources", "naming", "significance" and "wording"), and which action was intended (requesting or providing help, requesting an edit, announcing an edit that the user had made, criticizing the article without a direct request for action, coordinating actions between users, or referring to an established Wikipedia policy).

Regarding comments focused on the activity of editing, the "results were consistent with previous research, with a high frequency of discussion contributions about content and layout". The authors found that "the Japanese Wikipedia was the only one with more discussion contributions about layout than content when the discussion was about editing activities (40.18%)" and speculate that this is because "in the older, or larger, Wikipedias, practices and policies are likely to be better established than in the younger, or smaller, Wikipedias leading to a lower frequency of discussions about layout." (However, they later point out that the Finnish Wikipedia, rather than the Japanese, is the smallest and youngest among the three examined ones, noting that it shows a much higher frequency of discussion about policy—15.0%, versus 6.0% on the French and 3.3% on the Japanese Wikipedia.) In this class of comments, "discussions about citing sources were relatively common in the Finnish and French Wikipedias (18.8% and 12.4%, respectively). In the Japanese Wikipedia, sources were less common with 7.1% of all discussion contributions regarding editing activities."

Most discussions about translation activities were about *naming*—that is, "resolving the proper form for the title of the article, section or sub-section, names or proper nouns, and transliteration in the corresponding article",

contrasting the researchers' initial hypothesis that such discussion would "have a high frequency of contributions regarding translation of specific words and expressions" (their "naming" category "does not include phrasing or resolving proper translation of individual words or expressions"). As one reason, they identify "the diversity in naming practices of events between different language sources, such as mass media. Especially in the Finnish Wikipedia, discussion about sources was common (16.15%). These two topics are loosely related, as direct translations of the names of well-known events are often not acceptable in the target language Wikipedia."

Having identified naming issues and the search for suitable sources in the target language as "key problems" emerging in the translation discussions, the authors conclude that "the current approaches for supporting Wikipedia translation are not necessarily solving the main problems in Wikipedia translation" and proceed to suggest two "directions for designing supporting tools for Wikipedia translation, especially through open source development of MediaWiki extensions":

- "**Support for consistent translation of names and proper nouns**", e.g. by making a "user editable multilingual dictionary resource" directly accessible in the design, and enabling editors to "coordinate through discussion pages directly related to a specific dictionary or dictionary entry in order to resolve inconsistencies in a centralized repository"
- "**Support for citing sources in translated articles**", by offering an automatic search for sources that have themselves been translated into the target language and/or the development of a supporting "crowdsourcing translation tool for open content sources not available in the target language using machine translation"

The paper makes references to previous work on Wikipedia translation (including the authors' own), but does not mention the EU-supported CoSyne project [7], which aims to integrate tools with MediaWiki that "automate the dynamic multilingual synchronization process of Wikis" and would seem to have a lot of overlap with the kind of tools discussed in the paper.

New algorithm provides better revert detection

A paper[8] by three researchers affiliated with the EU-supported RENDER project (to be presented at next month's "Hypertext 2012" conference) promises "accurate revert detection in Wikipedia". The article starts by describing the detection of reverts as "a foundational step for many (more elaborated) research ideas, [whose] purposeful handling leads to a superior understanding of wiki-like systems of collaboration in general", giving an overview over such research. (Revert detection has also been used in tools for the use of the editing community, such as this one [9] that identify articles on the German Wikipedia that are currently controversial.)

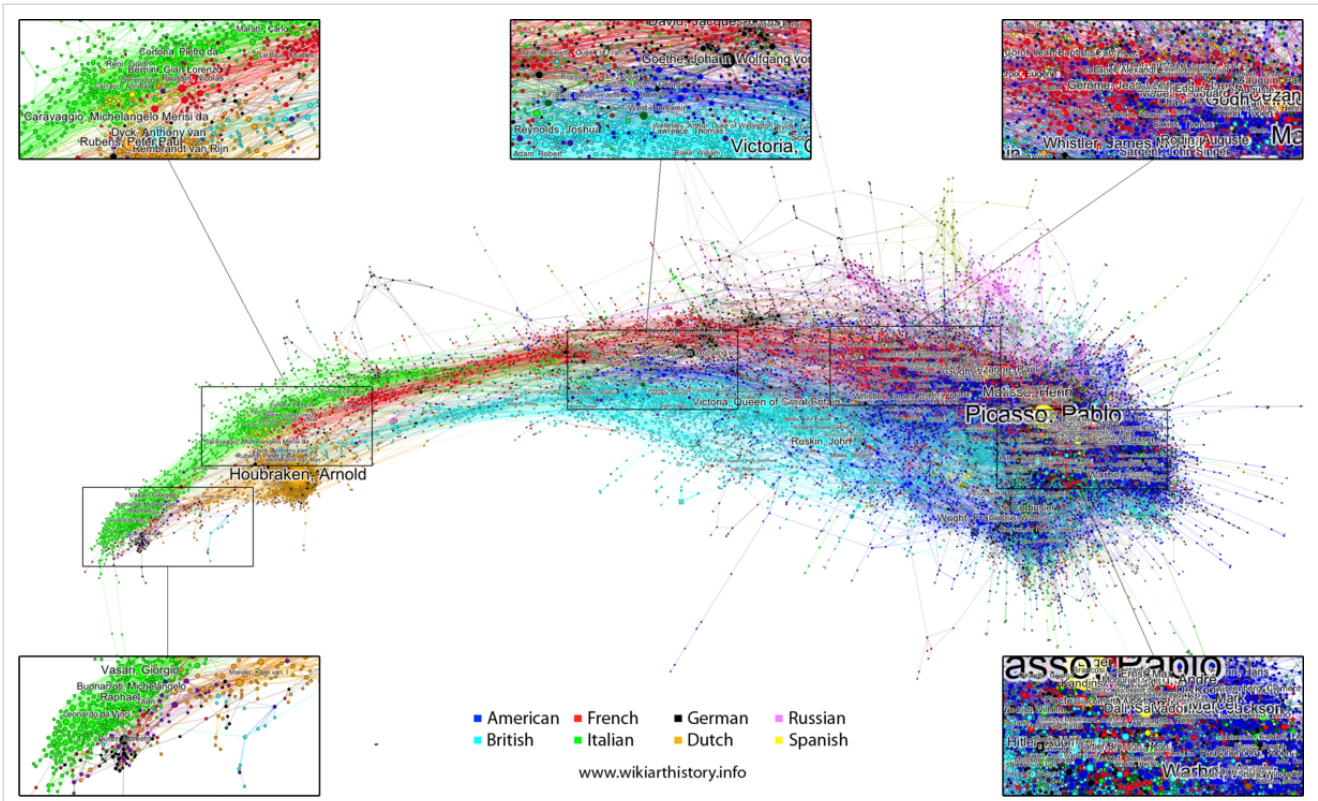
Overviewing the "state-of-the-art in revert detection", the authors criticize the prevalent "identity revert detection method" (SIRD) which relies on finding identical revisions using MD5 hashes, arguing that it does not fully match the definition of a revert in the (English) Wikipedia's policies at Wikipedia:Reverting: The SIRD method "does not require the reverting edit to actually undo the actions of an edit identified as reverted ... [Furthermore, it] is not possible to indicate if the reverting edit fully, partly or not at all undid the actions of the reverted edit ... It also does not require the intention of the reverting edit to revert any other edit." (Still, mainly due to requests by researchers [10], MD5 hashes have been integrated directly into the revision table stored by MediaWiki recently, necessitating considerable technical efforts when updating the existing databases for Wikimedia projects.)

The paper then presents the authors' new method for revert detection, which still aims to detect full reverts and to avoid false positives, while coming closer to the Wikipedia community's definition. It is implemented as an algorithm based on splitting the revisions' wikitext into word tokens (and made available online as a Python script [11]). Also, MD5 hashes are still used on a paragraph level to be able to detect unchanged paragraphs easily and speed up computation. The algorithm was then evaluated by a panel of Wikipedians recruited on the English Wikipedia in comparison with the existing SIRD method.

As summarized by the authors, this user study found the new method to be "more accurate in identifying full reverts as understood by Wikipedia editors. More importantly, our method detects significantly fewer false positives than

the SIRD method [27% in the sample, which however was somewhat small]". As a drawback, the authors note "the increased computational cost. As [the new algorithm] is quadratic over the number of words in the DIFFs [the changed text between subsequent revisions], in its current implementation it might not be the tool of choice if larger amounts of articles are to be analyzed; especially in the case of complete history dumps of the large Wikipedias, e.g., English, German or Spanish."

Briefly



The history of art mapped using Wikipedia (visualization of wikilinks between "art-historical actors" spanning at most 75 years, from Goldfarb et al.)

- **The history of art mapped using Wikipedia**: A paper by four researchers from Vienna, to be presented at next month's WebSci 2012 conference [12] examines the wikilinks between 18,002 Wikipedia articles about artists (or more precisely "art-historical actors", derived from the English Wikipedia via DBpedia), from present times back to ancient Greece. A first result appears to confirm the assumption that artists are more likely to be influenced by or related to their contemporaries: "the number of *short* links covering 0–37.5 years clearly outnumbers the sum of all the other .... This can be interpreted as such that contemporaries are much more likely to be interlinked than persons who are generations apart" They present a visualization [13] of the link graph colored by nationality of the person, which "reveals interesting patterns of cultural interaction within the network, as they are perceived by the English speaking Wikipedia community: The left side ... is dominated by Italians (green). This cluster spans Renaissance and Baroque times, fading out by the end of the 17th century. A small cluster on the lower left represents German Renaissance around Albrecht Duerer (black) ... The rightmost part represents Post-Modernist Americans, with a nationality-independent cluster of Architects beneath."
- **The use of references in Wikipedia coverage of current events**: On the blog of Ushahidi[14], Wikimedia researcher Heather Ford described preliminary "key findings" from an ongoing project examining the use of sources in Wikipedians work on current events such as the 2011 Egyptian Revolution: "1. The source <original version of the article and its author> of the page can play a significant role 2. Primary sources are gradually replaced by secondary sources," 3. The cite is not always the same as the source ("the citation that editors use to back up a particular phrase are not always the same as the source from which they receive their information"), 4.

The blurring of boundaries along traditional “reliable sources” lines. Her "design recommendations include the design of source management systems around the kind of collaboration that is already working on Wikipedia: where editors collaborate around specific news stories, checking to see whether the source actually reflects the information in the article, whether the source is accurately contextualized, whether other media verify the facts in the article and whether there is any accompanying multimedia."

- **Distribution of article title lengths**: A statistical analysis of the length of the more than 40 million article titles on all Wikipedias (including redirects) found that 90% are shorter than 32 characters and 98% are shorter than 53 characters. The blog post[15] by Denny Vrandečić, head of the Wikidata development team—who generated those stats to inform some design decisions for this project—provides charts of the length distribution for each language, exhibiting some interesting differences and similarities (e.g. the distributions for the English, German, French, Polish and Russian Wikipedias, as well as the overall one, peaks around 13 charaters).
- **To understand a Wikipedia article, which others does one need to read first?**: A paper titled "Crowdsourced Comprehension: Predicting Prerequisite Structure in Wikipedia[16] starts from the assumption that "the primary reason that technical documents are difficult to understand is lack of modularity: unlike a self-contained document written for a general reader, technical documents require certain background knowledge to comprehend—while that background knowledge may also be available in other on-line documents, determining the proper sequence of documents that a particular reader should study is difficult". Trying to develop a method to solve this problem in the example of five Wikipedia articles (global warming, meiosis, Newton's laws of motion, parallel postulate and public-key cryptography), the researchers analyzed the structure of wikilinks, whether pages had been edited by the same users, and the page text itself, and had Mechanical Turk workers decide in advance for many pairs of linked articles (within a subject domain) whether one was a prerequisite to understand the other. They conclude that "while it is not immediately obvious that this task is feasible, our experiments suggest that relatively reliable features to predict prerequisite structure exist, and can be successfully combined using standard machine learning methods".
- **High-conflict areas may deter uninvolved users**: A student thesis from Macalester College, titled "Characterizing Conflict in Wikipedia" [17] examines editing disputes between Wikipedians that concern several articles, pointing out that much of the previous research has only looked at such conflicts one article at a time. The analysis involved clustering 1.4 million articles. Among the conclusions is that "The vast majority of conflicts are very small, but there are still thousands of conflicts involving at least one hundred users. Conflicts between small numbers of users, or with small numbers of reverts, tend to span only one article, whereas larger conflicts tend to span more than one article." Also, within a conflict cluster, "contributions from users uninvolved in conflicts are even lower than those involved in conflicts. This indicates that users may be deterred from contributing to areas with high concentrations of conflict."
- **The vandalism revert and other temporal motifs, and their change from 2001 to 2011**: A paper presented at ICWSM '12[18] looks—like several other recent papers—at the bipartite graph of editors and the articles they have edited, but enriches it "with temporal information of both who edited the article [discerning bots, IP editors, and admins], and how the article was changed [. This] enables discovering meaningful editing behavior in the form of network motifs. These temporal motifs are repeated subgraphs of the editing graph which correspond to significant patterns of collaborative interactions." (The concept of network motifs is popular in bioinformatics, where it is applied to gene regulatory networks. See also the review of a earlier paper applying a simpler kind of motifs to analyze the editors-articles graph: "Collaboration pattern analysis: Editor experience more important than 'many eyes'".) Motifs involving just a single author were the most frequent. As an example of the patterns that become visible by including temporal information, among the multi-author motifs those involving a revert "occur much faster, with 6,558 of all 13,961 such motifs having a median time under 5 minutes ... The strong correspondence between reverting an edit and combating vandalism suggests that such short durations are due to active participation by Wikipedia community members, such as the Counter Vandalism Unit, which actively monitors recent revisions for potential vandalism". The authors then look at how the frequency of their motifs has

changed over the history of Wikipedia from 2001 to 2011, and find that "the trends suggest that the early growth was fueled by content addition from single authors or collaborating between two authors (B) and contributions from administrators. These early behaviors have given way to increases in behaviors associated with editing (A) and maintaining quality or vandalism detection (D)."

- **The Wikipedia research behind Google's new Knowledge Graph?**: On May 16, Google introduced its Knowledge Graph, a semantic network drawing information from many different sources including Wikipedia, which Google uses to enhance its search engine results with semantic information—often appearing to include excerpts from the infobox and lead section of a particular Wikipedia article on the top right corner of the results page. Two days later, Google Research announced [19] a paper by two Google employees titled "A Cross-Lingual Dictionary for English Wikipedia Concepts"[20] describing the construction of "a resource for automatically associating strings of text [such as search terms] with English Wikipedia concepts", considering "each individual Wikipedia article as representing a concept (an entity or an idea), identified by its URL". The resulting dataset is available for download [21] and described as having been "designed for recall [rather than precision]. It is large and noisy, incorporating 297,073,139 distinct string-concept pairs, aggregated over 3,152,091,432 individual links".

References

[1] Cimini, N., & Burr, J. (2012). An Aesthetic for Deliberating Online: Thinking Through “Universal Pragmatics” and “Dialogism” with Reference to Wikipedia. The Information Society, 28(3), 151–160. Routledge. DOI:10.1080/01972243.2012.669448 Closed access

[2] Gurunath Kulkarni, R., Trivedi, G., Suresh, T., Wen, M., Zheng, Z., & Rose, C. (2012). Supporting collaboration in Wikipedia between language communities. Proceedings of the 4th international conference on Intercultural Collaboration – ICIC ’12 (p. 47). New York, New York, USA: ACM Press. DOI:10.1145/2160881.2160890 Closed access

[3] Tanner, R. (2012). *Creating a Semantic Graph from Wikipedia*. Computer Science Honors Theses. Paper 29. http://digitalcommons.trinity.edu/compsci_honors/29/Open access

[4] http://simple.wikipedia.org

[5] Collier, B., & Kraut, R. (2012). Leading the Collective: Social Capital and the Development of Leaders in Core–Periphery Organizations. Physics and Society. http://arxiv.org/abs/1204.3682 Open access

[6] Gurunath Kulkarni, R., Trivedi, G., Suresh, T., Wen, M., Zheng, Z., & Rose, C. (2012). Supporting collaboration in Wikipedia between language communities. Proceedings of the 4th international conference on Intercultural Collaboration – ICIC ’12 (p. 47). New York, New York, USA: ACM Press. DOI:10.1145/2160881.2160890 Closed access

[7] http://www.cosyne.eu/wiki/Work_Packages

[8] Fabian Flöck, Denny Vrandecic and Elena Simperl. Reverts Revisited – Accurate Revert Detection in Wikipedia (http://people.aifb.kit.edu/ffl/reverts/ht083-floeckATS.pdf). HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA. Open access

[9] https://toolserver.org/~aka/cgi-bin/revstat.cgi

[10] http://www.gossamer-threads.com/lists/wiki/wikitech/245939

[11] http://people.aifb.kit.edu/ffl/reverts/

[12] Doron Goldfarb, Max Arends, Josef Froschauer, Dieter Merkl. Art History on Wikipedia, a Macroscopic Observation (http://vsem.ec.tuwien.ac.at/wp-content/publications/WEBSCI12.pdf) WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA. Open access

[13] http://vsem.ec.tuwien.ac.at/wikiarthistory/

[14] Ford, Heather: Update on the Wikipedia sources project (http://blog.ushahidi.com/index.php/2012/05/17/update_on_wikisources_project/). Ushahidi.com, May 17, 2012 Open access

[15] Vrandečić, D. (2012). Distribution of title lengths in Wikipedias (http://simia.net/wikititles/). simia.net, 10 May 2012 Open access

[16] Talukdar, P. P., & Cohen, W. W. (2012). *Crowdsourced Comprehension: Predicting Prerequisite Structure in Wikipedia*. 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL 2012. **PDF** (http://www.cs.cmu.edu/~wcohen/postscript/naacl-ws-2012-partha.pdf) Open access

[17] Miller, N. (2012). *Characterizing Conflict in Wikipedia*. Honors Projects. Paper 25. http://digitalcommons.macalester.edu/mathcs_honors/25 Open access

[18] Jurgens, D., & Lu, T.-ching. (2012). Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia. *ICWSM '12* **PDF** (http://www.cs.ucla.edu/~jurgens/papers/jurgens-lu-icwsm-2012.pdf) Open access

[19] http://googleresearch.blogspot.com/2012/05/from-words-to-concepts-and-back.html

[20] Valentin I. Spitkovsky, Angel X. Chang. " A Cross-Lingual Dictionary for English Wikipedia Concepts (http://nlp.stanford.edu/pubs/crosswikis.pdf)". Eighth International Conference on Language Resources and Evaluation (LREC 2012) Open access

[21] http://www-nlp.stanford.edu/pubs/crosswikis-data.tar.bz2/

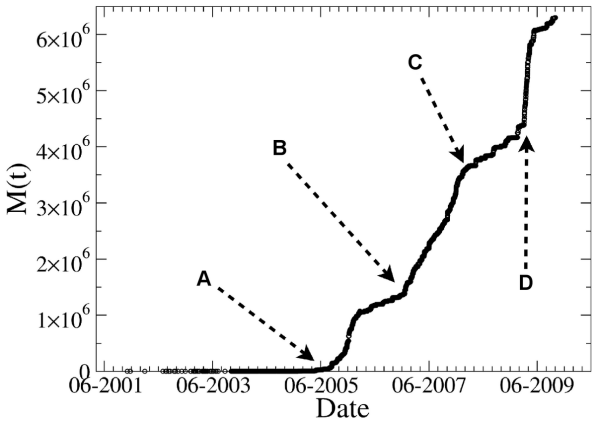
Issue 2(6): June 2012

Edit war patterns, deleters vs. the 1%, never used cleanup tags, authorship inequality, higher quality from central users, and mapping the wikimediasphere

With contributions by: Tbayer, Piotrus, Evan and Daniel Mietchen

Dynamics of edit wars

"Dynamics of Conflicts in Wikipedia"[1], develops an interesting "measure of controversiality", something that might be of interest to editors at large if it was a more widely popularized and dynamically updated statistic. The authors look at the patterns of edit warring on Wikipedia articles, finding that edit warriors are usually prone to reaching consensus, and the rare cases of never-ending warring involve those that continuously attract new editors who have not yet joined the consensus.



Controversy about Michael Jackson as quantified on the basis of reverted edits to his Wikipedia article. A: Jackson is acquitted on all counts after five month trial. B: Jackson makes his first public appearance since the trial to accept eight records from the Guinness World Records in London, including Most Successful Entertainer of All Time. C: Jackson issues *Thriller 25*. D: Jackson dies in LA.

Regarding methodology, the authors' decision to filter out articles with under 100 edits as "evidently conflict-free" is a bit problematic, as articles with fewer than 100 edits have been subject to clear, if not over-long, edit warring (a recent example: Concerns and controversies related to UEFA Euro 2012). One could also wish that the discussion of the "memory effects" – a term mentioned only in the abstract and lead, which the author suggests is significant to understanding the conflict dynamic – was explained somewhere in the article (the term "memory" itself appears four times in the body and does not seem to be operationalized anywhere).

A press release accompanying the paper is titled "Wikipedia 'edit wars' show dynamics of conflict emergence and resolution [2]", while an MSNBC tech news headline summarized it as "Wikipedia is editorial warzone, says study [3]".

Who deletes Wikipedia

In a recent blog post [4] by Wibidata, an analytics startup based in San Francisco, the authors set out to shed light on the often-quoted claim that most of Wikipedia was written by a small number of editors, noting other editorial patterns along the way. Using the entirety of English Wikipedia revision history (they wanted to show that their platform can scale), the authors looked at the distribution of edits across editor cohorts, grouped by number of total edits. They found that from a pure count perspective, the most active 1% of editors had contributed over 50% of the total edits. (see original plot here [5])

In response to the suggestion that the strongly skewed distribution of edits might just be due to a core set of editors who primarily make only minor formatting modifications, they looked at the net number of characters contributed by each editor. Grouping editors by total number of edits as before, they showed an even more strongly skewed distribution, with the top 1% contributing well over 100% of the total number characters on Wikipedia (i.e. an amount of text that is larger than the current Wikipedia) and the bottom 95% of editors deleting more on average than they contributed (original plot [6]). Next, the authors separated logged in users from non-logged in "users"

(identified only by IP addresses) and recomputed the distribution of net character contributions. By edit-count cohort, logged-in users tended to contribute significantly more than their anonymous counterparts, and non-logged-in users tended to delete significantly more (original plot [7]).

In summary, low-activity and new editors, along with anonymous users, tend to delete more than they contribute; this reinforces the notion that Wikipedia is largely the product of a small number of core editors.

Evaluating and predicting interlingual links in Wikipedia

Published in proceedings of *SEM, a computational semantics conference, researchers from the University of North Texas and Ohio University looked into the nature of interlingual links on Wikipedia, both reviewing the quality of existing links and exploring possibilities for automatic link discovery.[8] The researchers took the directed graph of interlingual links on Wikipedia and used the lens of set-theoretic operations to structure an evaluation of existing links, to build a system for automatic link creation. For example, they suggest that the properties of symmetry and transitivity should hold for the relation of interlingual linking. This means that if there is an interlingual link from language A to B, there should also be a link from B to A, and if there is a link from language A to B, and language B to C, then there should be a link from language A to C. (This assumption is routinely made by the many existing Interwiki bots.) They further refine the notion of transitivity, by grouping article pairs by the number of transitive 'hops' required to connect a candidate article pair.

Their methodology revolves around the creation of a sizeable annotated gold data set. Using these labels, they first evaluated the quality of existing links, finding between one half and one third to fail their criteria for legitimate translations. They then evaluated the quality of various implied links. For example, reverse links where they do not already exist satisfy their criteria for faithful translation only 68% of the time.

The gold data set was used to train a boosted decision-tree classifier for selecting good candidate pairs of articles. They used various network topology features to encode the information in interlingual links for a given topic and found that they can significantly beat the baseline, which uses only the presence of direct links (73.97% compared with 69.35% accuracy).

"Wikipedia Academy" preview

Various conference papers and posters from the upcoming "Wikipedia Academy" (hosted by the German Wikimedia chapter from June 29 to July 1 in Berlin) are already available online [9]. A brief overview of those which are presenting new research about Wikipedia:

- **{{Citation needed}} more effective than {{unreferenced}}**: "On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia"[10] shows "that inline tags are more effective than tag boxes" in tagging article flaws so that they get remedied. The researchers also "reveal five cleanup tags that have not been used at all, and 15 cleanup tags that have been used less than once per year", recommending their deletion, and "ten cleanup tags that have been used, but the tagged flaws have never been fixed." Similar to a paper reviewed in the April issue of this report (One in four of articles tagged as flawed, most often for verifiability issues"), they find that "the majority (71.62%) of the tagged articles have been tagged with a flaw that belongs to the flaw type *Verifiability*".
- **A paper titled "The Power of Wikipedia: Legitimacy and Territorial Control"**[11], is "based on the experience of the projects WikiAfrica (2006-2012) and Share Your Knowledge (2011-2012)", and looks at various aspects of Wikipedia, Wikimedia chapters and the Foundation through the lens of "anthropological, african and post-colonial studies."
- **"Individual and Cultural Memories on Wikipedia and Wikia, Comparative Analysis"**[12] looks at the coverage of the late British DJ John Peel on Wikipedia and Wikia, respectively, as well as the Wikipedia article about the 1980s.

- An "Extended Abstract", **"Latent Barriers in Wiki-based Collaborative Writing"**[13] compares the collaborative process "25 special-purpose wikis" (most of them hosted by Wikia) with that of the German Wikipedia. One observation of the work in progress is a "strong divide between extracts of Wikipedia (even if being reduced to single articles and their one-link neighborhoods) on the one hand and special purpose wikis on the other."
- Two Brazilian authors will examine **"the climate change controversy through 15 articles of Portuguese Wikipedia"**.[14] The paper contains various quantitative results about the edit history of these articles, some of them unsurprising ("A very strong positive correlation (0.994) was found between the number of edits and the number of editors of an article"). Using the framework of actor–network theory, the authors conclude that "the collaborative encyclopedia is enrolled as an ally for the mainstream science and becomes one of its spokespersons."
- **Historical infobox data**: An article by four authors from Google Switzerland and the Spanish National University of Distance Education (UNED) observes[15] that "much research has been devoted to automatically building lexical resources, taxonomies, parallel corpora and structured knowledge from [Wikipedia]", often using the structured data present in infoboxes (which they say are present in "roughly half" of English Wikipedia articles). However, this research has so far used only snapshots representing the state of articles at a particular point in time, whereas their project embarked to extract "a wealth of historical information about the last decade ... encoded in its revision history." The resulting 5.5GB dataset, called "Wikipedia Historical Attributes Data (WHAD)", will be made freely available for download.
- **Better authorship detection, and measuring inequality**: Two researchers from the University of Karlsruhe will present an algorithm[16] to detect which user wrote which part of a Wikipedia article. Similar to a new revert-detection algorithm presented in a recent paper co-authored by one of the present authors (see last month's issue: "New algorithm provides better revert detection"), one crucial part of the algorithm is to split the article's wikitext into paragraphs, analyzing them separately under the assumption "that most edits (if they are not vandalistic) change only a very minor part of an article's content". Another part is calculating the cosine similarity of sentences that are not exactly identical. In the authors' own test, the new algorithm performed significantly better than the widely used WikiTrust/WikiPraise tool. Having determined the list of authors for an article revision and the size of each author's contribution, they then define a gini coefficient "as an inequality measure of authorship" (roughly, an article written by a single author will have coefficient 1, while one with equal contributions by a multitude of editors will have coefficient 0). They implement a tool called "WIKIGINI" to plot this coefficient over an article's history, and show a few examples to demonstrate that it "may help to spot crucial events in the past evolution of an article". The paper starts out from the assumption "that the concentration of words to just a few authors can be an indicator for a lack of quality and/or neutrality in an article", but it does not (yet) contain a systematic attempt to correlate the gini coefficient and existing measures of article quality.
- **Troll research compared**: A paper by a German Wikipedian titled "Here be Trolls: Motives, mechanisms and mythology of othering in the German Wikipedia community"[17] examines four academic texts about online trolls (only one of them in the context of Wikipedia), which "were compared regarding their scope, their theoretical approach, their methods and their findings concerning trolls and trolling.

Posters

- **"Self-organization and emergence in peer production: editing 'Biographies of living persons' in Portuguese Wikipedia**[18]
- **"Biographical articles on Serbian Wikipedia and application of the extraction information on them"**[19]
- **"Wikipedia article namespace – user interface now and a rhizomatic alternative"**[20]
- **"Extensive Survey to Readers and Writers of Catalan Wikipedia: Use, Promotion, Perception and Motivation"**[21]

Researcher Felipe Ortega blogged[22] about a new parser for Wikipedia dumps, to be integrated into "WikiDAT (Wikipedia Data Analysis Toolkit) ... a new integrated framework to facilitate the analysis of Wikipedia data using

Python, MySQL and R. Following the pragmatic paradigm 'avoid reinventing the wheel', WikiDAT integrates some of the most efficient approaches for Wikipedia data analysis found in libre software code up to now", which will be featured in a workshop at the conference.

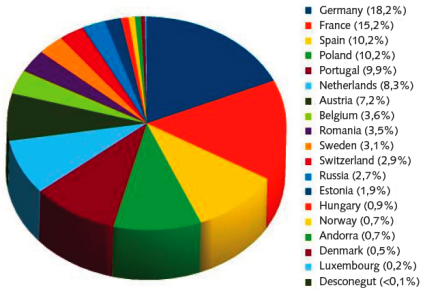
Special issue of "Digithum" on Wikipedia research

The open-access journal "Digithum" (subtitled "The Humanities in the Digital Era") has published a special issue [23] containing five papers about Wikipedia from various disciplines, with a multilingual emphasis (including research about non-English Wikipedias, and Catalan and Spanish versions of the papers alongside the English versions):

- **Are articles about companies too negative?**: A paper titled "Wikipedia's Role in Reputation Management: An Analysis of the Best and Worst Companies in the United States"[24] looked at the English Wikipedia articles about the ten companies with the best and worst reputations according to the "Harris Reputation Quotient", a 2010 online survey about "perceptions for 60 of the most visible companies in America". Those 20 articles were coded, sentence by sentence, as positive, negative or neutral, and according to other "reputation attributes". Among the findings was that "the companies with the worst reputations had more negative content; they had, in fact, almost double the amount of negative content, although only slightly less positive content. Both types of companies had more negative than positive content. This indicates that even if a company is considered to have a good reputation, it is still very vulnerable to having its dirty laundry aired on Wikipedia." Another observation was that "emotional appeal is an attribute where both types of companies lacked content. It was rare for companies to have content about trust or feeling good, which only existed for the best companies" (an interesting question may be whether this is related to Wikipedia guidelines such as WP:PEACOCK). The paper appears at a time where many PR industry professionals in the US and UK argue that Wikipedia should allow them more control over the articles about their clients, and ends by highlighting the "importance of public relations professionals monitoring and requesting updates to Wikipedia articles about their companies". This conclusion resembles that of another recent study by one of the authors (DiStaso), which likewise concerned company articles, implicating a somewhat controversial conclusion about their accuracy (see the April issue of this research report: "Wikipedia in the eyes of PR professionals").
- **WordNets from Wikipedia**': The second paper[25] describes "the state of the art in the use of Wikipedia for natural language processing tasks", including the researchers' own application of Wikipedia to build WordNet databases in Catalan and Spanish.
- **The Wikimedia movement as "wikimediasphere"**: The article "Panorama of the wikimediasphere"[26] gives an overview of the Wikimedia movement, proposing the term "wikimediasphere" to describe it, and explaining "the role of the communities of editors of each project and their autonomy with respect to each other and to the Wikimedia Foundation", which is seen as "the principal supplier of the technological infrastructure and also the principal instrument for obtaining economic and organisational resources". Its vision statement is presented as a summary of the aim that is "the ideological glue that binds all the players involved". The section about "the social and institutional dimension" of the sphere briefly covers the Foundation's governance and funding models, Wikimedia chapters and other recognized supporting organizations, and the various wikis and other online platforms that structure "the organisational activity": The Foundation wiki, Meta-wiki, Strategy wiki, Outreach wiki, the Wikimedia blog and the blogs of community members aggregated on Planet Wikimedia, mailing lists etc. Authored by a Wikimedian who is a member of both the Spanish chapter and the Catalan "Friends of Wikipedia" association, the paper is remarkably well-informed and up-to-date, e.g. incorporating the Board resolution on "Recognized Models of Affiliations" from the beginning of April, and various other recent events such as the English Wikipedia's SOPA/PIPA blackout. The abstract uses the term "WikiProjects" in a different sense from that common among English-speaking Wikimedians, possibly a translation error.
- **Truth and NPOV**: The fourth article[27] by Nathanial Tkacz (one of the organizers of the "Critical Point of View"/CPOV initiative that organized three conferences about Wikipedia in 2010, see *Signpost* interview) sets out to "show that Wikipedia has in fact two distinct relations to truth: one which is well known and forms the

basis of existing popular and scholarly commentaries, and another which refers to equally well-known aspects of Wikipedia, but has not been understood in terms of truth. I demonstrate Wikipedia's dual relation to truth through a close analysis of the Neutral Point of View core content policy (and one of the project's 'Five Pillars')."

- **Wiki Loves Monuments**: A paper titled "Wiki Loves Monuments 2011: the experience in Spain and reflections regarding the diffusion of cultural heritage",[28] written by five Spanish Wikimedians, gives a concise overview of the photo contest as it played out in Spain last year.



Wiki Loves Monuments 2011 uploads by country

Briefly

- **Who was notable in London in the 1960s?**: A Master's thesis in Computer Science[30] describes "A tool for extracting and indexing spatio-temporal information from biographical articles in Wikipedia". The tool, named "Kivrin" after a time-travelling character from a science fiction novel, is available online [31], and grew out of an earlier, simpler one that searches for articles about plants and animals living at a particular geographical place ("Flora & Fauna Finder [32]"). The author remarks that "the data is skewed, like Wikipedia itself, towards the U.S. and Western Europe and relatively recent history". A search for the 1960s in London [33] brings up several Beatles-related biographies near the top. While the tool does seem to cover languages other than English (e.g. text from the Hungarian entry on Gottlob Frege appears in the search results [34] for Jena, the German town), searches for Hungarian or other non-English place names (e.g. Moszkva [35] and Москва [36], the Hungarian and Russian names of Moscow) yielded no results. Disambiguation is attempted by way of geocodes but far from robust - the search results [37] for Halle, Saxony-Anhalt actually contain multiple entries referring to Halle, North Rhine-Westphalia.


Bill Harry (pictured with his wife Victoria) is one of the articles found by Kivrin for "London in the 1960s".

- **How did people in Europe feel in the 1940s?**: As described in a post in the New York Times' "Bits" blog[38] Kalev Leetaru from the University of Illinois conducted a sentiment analysis of statements on Wikipedia connected to a particular space and time, and made the result into a video: "The Sentiment of the World Throughout History Through Wikipedia [39]"


The bonobo (here a juvenile) is amongst the species that the Flora and Fauna finder finds [29] for Congo.

- **One third of the average Wikipedia consists of interwiki links**: According to an analysis[40] by Denny Vrandečić, head of the Wikidata development team, "on average, 33% of a Wikipedia is language links. In total, there are 240 Mio of them, 5GB [41]" (making up 5.3% of the overall text across all languages). The ratio tends to be higher on smaller Wikipedias.
- **"Central" users produce higher quality**: A preprint by two Dublin-based researchers attempts "Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality".[42] The former uses the

assumption that contributions which survive many subsequent edits tend to have a higher quality, and "measures the quality of an article by aggregating the edit longevity of all its author contributions". The second approach considers either the coauthorship network (the bipartite graph of users and the articles they have edited, used in many recent papers to grasp Wikipedia's collaboration processes) or the user talk page (UTP) network, where two Wikipedians are connected if one has edited the other's talk page. It is assumed that a user's "centrality" in one of these networks is a measure for the "contributor authoritativeness". These quality measures are then evaluated on 9290 history-related Wikipedia articles against the manual quality rating from WikiProject History. "The results suggest that it is useful to take into account the contributor authoritativeness (i.e., the centrality metrics of the contributors in the Wikipedia networks) when assessing the information quality of Wikipedia content. The implication for this is that articles with significant contributions from authoritative contributors are likely to be of high quality, and that high-quality articles generally involve more communication and interaction between contributors."

- **Familiarity breeds trust**: A bachelor thesis[43] at Twente University had 40 college students assess the trustworthiness of articles from the English Wikipedia, after a search for a piece of information in the article that was either present at the top or near the bottom of the article. The hypothesis that the longer search in the second case might affect the trustworthiness rating was rejected by the results, but it was found (consistent with other research) that "Trust was higher in articles with a familiar topic, rather than with unfamiliar topics".

References

[1] Yasseri, Taha; Sumi, Robert; Rung, András; Kornai, András; Kertész, János (2012). Szolnoki, Attila. ed. "Dynamics of Conflicts in Wikipedia". PLoS ONE 7 (6): e38869. DOI:10.1371/journal.pone.0038869 PDF (http://dx.plos.org/10.1371/journal.pone.0038869) Open access, preprint

[2] http://phys.org/news/2012-06-wikipedia-wars-dynamics-conflict-emergence.html

[3] http://www.technolog.msnbc.msn.com/technology/technolog/wikipedia-editorial-warzone-says-study-838793

[4] Who Deletes Wikipedia (http://www.wibidata.com/2012/06/06/who-deletes-wikipedia)

[5] http://www.wibidata.com/wp-content/uploads/2012/06/PercentOfRevisionsByEdGroup.png

[6] http://www.wibidata.com/wp-content/uploads/2012/04/PercentOfDeltaByEditingActivity.png

[7] http://www.wibidata.com/wp-content/uploads/2012/04/AveDeltaByEditActivity.png

[8] Dandala, B., Mihalcea, R., & Bunescu, R. (n.d.). Towards Building a Multilingual Semantic Network: Identifying Interlingual Links in Wikipedia. Retrieved from PDF (http://ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL004.pdf)

[9] http://wikipedia-academy.de/2012/wiki/Accepted_Submissions

[10] Maik Anderka, Benno Stein and Matthias Busse: On the Evolution of Quality Flaws and the Effectiveness of Cleanup Tags in the English Wikipedia (PDF) (http://wikipedia-academy.de/2012/w/images/f/f0/13_Paper_Maik_Anderka_Benno_Stein_Matthias_Busse.pdf) Open access

[11] Iolanda Pensa: The Power of Wikipedia: Legitimacy and Territorial Control (PDF) (http://wikipedia-academy.de/2012/w/images/8/8f/17_Paper_Iolanda_Pensa.pdf) Open access

[12] Simeona Petkova: Individual and Cultural Memories on Wikipedia and Wikia, Comparative Analysis (PDF) (http://wikipedia-academy.de/2012/w/images/2/2d/28_Paper_Simeona_Petkova.pdf) Open access

[13] Alexander Mehler, Christian Stegbauer and Rüdiger Gleim: Latent Barriers in Wiki-based Collaborative Writing (PDF) (http://wikipedia-academy.de/2012/w/images/8/8f/12_Paper_Alexander_Mehler_Christian_Stegbauer_RÃ¼diger_Gleim.pdf) Open access

[14] Bernardo Esteves and Henrique Cukierman: The climate change controversy through 15 articles of Portuguese Wikipedia (PDF) (http://wikipedia-academy.de/2012/w/images/c/c6/5_Paper_Bernardo_Esteves_Henrique_Cukierman.pdf) Open access

[15] Guillermo Garrido, Enrique Alfonseca, Jean-Yves Delort and Anselmo Peñas: " Extracting Wikipedia Historical Attributes Data (http://wikipedia-academy.de/2012/w/images/7/7c/31_Paper_Guillermo_Garrido_Enrique_Alfonesca_Jean-Yves_Delort_Anselmo_Penas.pdf)" (PDF) Open access

[16] Fabian Flöck and Andriy Rodchenko: Whose article is it anyway? – Detecting authorship distribution in Wikipedia articles over time with WIKIGINI (PDF) (http://wikipedia-academy.de/2012/w/images/2/24/23_Paper_Fabian_FlÃ¶ck_Andriy_Rodchenko.pdf) Open access

[17] Moritz Braun: Here be Trolls: Motives, mechanisms and mythology of othering in the German Wikipedia community (PDF) (http://wikipedia-academy.de/2012/w/images/3/37/20_Paper_Moritz_Braun.pdf) Open access

[18] Carlos D'Andréa: Seft-organization and emergence in peer production: editing “Biographies of living persons” in Portuguese Wikipedia (PDF) (http://wikipedia-academy.de/2012/w/images/9/9d/10_Poster_Carlos_D'Andrea.pdf) Open access

[19] Djordje Stakic: Biographical articles on Serbian Wikipedia and application of the extraction information on them (PDF) (http://wikipedia-academy.de/2012/w/images/8/8d/14_Poster_Djordje_Stakic.pdf) Open access

[20] Stephan Ligl: Wikipedia article namespace – user interface now and a rhizomatic alternative (PDF) (http://wikipedia-academy.de/2012/w/images/3/37/19_Poster_Stephan_Ligl.pdf)
[21] Marc Miquel-Ribé, David Morera-Ruíz and Joan Gomà-Ayats: Extensive Survey to Readers and Writers of Catalan Wikipedia: Use, Promotion, Perception and Motivation (PDF) (http://wikipedia-academy.de/2012/w/images/1/10/29_Poster_David_Morera-Ruiz_Marc_Miquel-RibÃ©_Joan_GomÃ -Ayats.pdf) Open access
[22] Ortega, Felipe: " Improving the extraction of Wikipedia data (http://libresoft.es/node/564)" libresoft.es, 2012-06-03
[23] http://digithum.uoc.edu/ojs/index.php/digithum/user/setLocale/en_US?source=%2Fojs%2Findex.php%2Fdigithum%2Fissue%2Fview%2Fn14
[24] Marcia W. DiStaso, Marcus Messner: " Wikipedia's Role in Reputation Management: An Analysis of the Best and Worst Companies in the United States (http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-distaso-messner)" DIGITHUM, NO 14 (2012) Open access
[25] Antoni Oliver, Salvador Climent: Using Wikipedia to develop language resources: WordNet 3.0 in Catalan and Spanish (http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-oliver-climent) Open access
[26] David Gómez Fontanills: " Panorama of the wikimediasphere (http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-gomez)" Open access
[27] Nathaniel Tkacz: " The Truth of Wikipedia (http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-tkacz)" Open access
[28] Emilio José Rodríguez Posada, Ángel González Berdasco, Jorge A. Sierra Canduela, Santiago Navarro Sanz, Tomás Saorín: Wiki Loves Monuments 2011: the experience in Spain and reflections regarding the diffusion of cultural heritage (http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-rodriguez-gonzalez-sierra-navarro-saorin). Digithum, no. 14 (May, 2012), p. 94. Open access
[29] http://linserv1.cims.nyu.edu:48866/cgi-bin/classResults.cgi?place=Congo
[30] Morton-Owens, E. G. (2012). A tool for extracting and indexing spatio-temporal information from biographical articles in Wikipedia. New York University. PDF (http://www.cs.nyu.edu/web/Research/MsTheses/owens_emily.pdf)
[31] http://linserv1.cims.nyu.edu:48866/cgi-bin/index.cgi
[32] http://linserv1.cims.nyu.edu:48866/cgi-bin/classSearch.cgi
[33] http://linserv1.cims.nyu.edu:48866/cgi-bin/results.cgi?toponym=London&startdate=1960&enddate=1969&latitude=51.50853&longitude=-0.12574
[34] http://linserv1.cims.nyu.edu:48866/cgi-bin/results.cgi?toponym=Jena&startdate=0&enddate=2012&latitude=50.93333&longitude=11.58333
[35] http://linserv1.cims.nyu.edu:48866/cgi-bin/toponym.cgi?startdate=0&enddate=2012&toponym=Moszkva&nameSearch=Submit+name
[36] http://linserv1.cims.nyu.edu:48866/cgi-bin/toponym.cgi?startdate=0&enddate=2012&toponym=%26%231052%3B%26%231086%3B%26%231089%3B%26%231082%3B%26%231074%3B%26%231072%3B&nameSearch=Submit+name
[37] http://linserv1.cims.nyu.edu:48866/cgi-bin/results.cgi?toponym=Halle&startdate=0&enddate=2012&latitude=51.5&longitude=12.0
[38] http://bits.blogs.nytimes.com/2012/06/14/how-big-data-sees-wikipedia/
[39] http://www.youtube.com/watch?v=KmCQVIVpzWg
[40] Vrandečić, D. (2012). Ratio of language links to full text in Wikipedias (http://simia.net/languagelinks/)" simia.net, June 2012
[41] https://twitter.com/vrandezo/status/217192948482322432
[42] Qin, X., & Cunningham, P. (2012). Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality. ArXiv, 15. Computers and Society, . Retrieved from http://arxiv.org/abs/1206.2517
[43] Hensel, T. (2012, March 11). Impact of duration of the search on the trust judgment of Wikipedia articles. Retrieved from http://essay.utwente.nl/61602/1/Hensel%2C_T.N.C.H._%2D_s0170860_(verslag).pdf

Issue 2(7): July 2012

Conflict dynamics, collaboration and emotions; digitization vs. copyright; WikiProject field notes; quality of medical articles; role of readers; Best Wiki Paper Award

With contributions by: Daniel Mietchen, Junkie.dolphin, Jodi.a.schneider, Adler.fa, OrenBochman, DarTar, Benjamin Mako Hill and Tbayer

Modeling social dynamics in a collaborative environment

A draft of a letter, submitted for publication, has been posted on ArXiv.[1] The letter reports research on modeling the process of collaborative editing in Wikipedia and similar open-collaboration writing projects. The work builds on previous research by some of its authors on conflict detection in Wikipedia. The authors explore a simple agent-based model of opinion dynamics, in which editors influence each other either by direct communication or by successively editing a shared medium, such as a Wikipedia page. According to the authors, the model, although highly idealized, exhibits a rich behavior that can reproduce, albeit only qualitatively, some key characteristics of conflicts over real-world Wikipedia pages. The authors show that, for a fixed editorial pool with one "mainstream" and two opposing "extremist" groups, consensus is always reached. However, depending on the values of the model's input parameters, achieving consensus may take an extremely long time, and the consensus does not always conform to the initial mainstream view. In the case of a dynamic group, where new editors replace existing ones, consensus may be achieved through a phase of conflict, depending on the rate of new editors joining the editorial pool and on the degree of controversy over the article's topic.

How Wikipedia articles benefit from the availability of public domain resources

In a copyright panel at this month's Wikimania, Abhishek Nagaraj – a PhD student and economist from the MIT Sloan School of Management – presented early results from an econometric study of copyright law. The study used data from the English Wikipedia's WikiProject Baseball to try to consider how gains from digitization are moderated by the effects of copyright. Previous work on the economics of copyrights have struggled to disentangle the effects of copyright with the effects of increased access that often coincides with content after it has entered the public domain.

The paper takes advantage of the fact that in 2008, Google digitized and published a large number of magazines as part of the Google Books projects. Among other magazines published were 70 years of back-issues (1945–1970) of *Baseball Digest*, a magazine that publishes baseball stories, statistics, and photographs. Measuring the effect of digitization, Nagaraj found that the articles on baseball All-Stars from between 1944 and 1984 saw large increases in size (5,200) around the period that the digital Google Books version of *Baseball Digest* became available. However, because of the law governing copyright expiration, all the issues of *Baseball Digest* published before 1964 were in the public domain, while issues published after were not. Using the econometric difference in differences technique, Nagaraj compared the different effects of digitization for (1) players who began their professional baseball career after 1964 and as a result had no new digitized public-domain material and (2) players who had played before and were thus more likely to have digitized material about them enter the public domain.

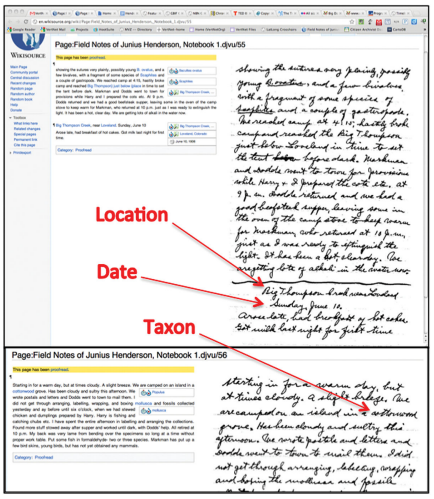
In terms of the effect of copyright, Nagaraj found no effect on the length of Wikipedia articles on public domain status but found a strong effect for images. Wikipedia writers could, presumably, simply rewrite copyrighted material or may not have found the baseball digest form appropriate for the encyclopedia. However, Nagaraj found that the availability of public domain material in baseball digest led to a strong increase in the number of images. Before Google Books published the material, the pre-64 group had an average of 0.183 pictures on their articles and the post 64 group had about 0.158 pictures. In the period after digitization, both groups increased but the older group increased more, to 1.15 pictures per article as opposed to 0.667 images for the more recent players whose *Baseball*

Digest material was still under copyright. Nagaraj also found that those players with public domain material have more traffic to their articles. The essay controls for a large number of variables related to players, their performance and talent, and their potential popularity, as well as for trends in Wikipedia editing.

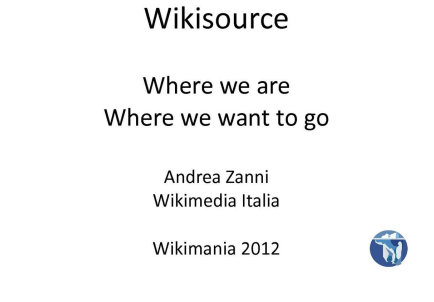
The presentation slides are available on the Wikimania conference website[2] and a nice journalistic write-up was published by The Atlantic [3].

Annotating field notes via Wikisource

Field notes can be a valuable source of information about meteorological, geological and ecological aspects of the past, and making them accessible by way of Wikisource-based semantic annotation was the focus of a recent study[4] published in *ZooKeys* as part of a special issue on the digitization of natural history collections. The paper described how the field notes of Junius Henderson from the years 1905–1910 have been transcribed on Wikisource and then semantically annotated, as illustrated in the screenshot. Henderson was an avid collector of molluscs and, while trained as a judge, served as the first curator of the University of Colorado Museum of Natural History. His notebooks are rich in species occurrence records, but also contain occasional gems like this one from September 3, 1905:



Extraction of location, date and taxon data from Field Notes of Junius Henderson on Wikisource



User:Aubrey's diagram of a future Wikisource, which combines text with additional layers of transcription, hypertext, annotations and comments.

> Train again so late as to afford ample opportunity for philosophic meditation upon the motives which inspire railroad people to advertise time which they do not expect to make except under rare circumstances

The article provides a detailed introduction to the workflows on the English Wikisource in general and to WikiProject Field Notes in particular, which is home to transcriptions of other field notes as well. The data resulting from annotation of the field notes are available [5] in Darwin Core format under a Creative Commons Public Domain Dedication (CC0). This work ties in with discussions that took place at Wikimania about the future of Wikisource, the technical prerequisites and existing tools and initiatives.

Quality of medical information in Wikipedia

The quality of medical information in Wikipedia could be vastly improved, based on the results of a recent study of 24 articles in pediatric otolaryngology[6] (more commonly referred to as "ear, nose, and throat" or ENT). The study compared results on common ENT diagnoses from Wikipedia, eMedicine, and MedlinePlus (the three most popular websites, by their determination) and they found that Wikipedia's articles on ENT were the least accurate and had the most errors of the three and that they were in the middle of the other two in regards to readability.

While one of the most referenced sources in this area, Wikipedia had poor content accuracy (46%) compared to the two other frequent sources. MedlinePlus has comparable (49%) accuracy, but was missing 7 topics. The clear leader in accuracy, eMedicine, suffers from a higher reading level. The study provides specific criteria, in section 2.3, which could be considered for evaluation of existing articles. One limitation of the study is that, while suggesting that Wikipedia "suffers from the lack of understanding that a physician-editor may offer", it does not point to information on how to get involved with Wikipedia. Engagement with the pediatric medicine community would be beneficial, especially since about 25% of parents made decisions about their children's care in part based on online information.

Emotions and dialogue

A forthcoming paper at this year's WikiSym conference investigates the emotions expressed in article and user talk pages[7]. "Administrators tend to be more positive than regular users", and the paper suggests that "as women gain experience in Wikipedia they tend to adopt the emotional tone of administrators", for instance linking to policy at more than twice the rate as males. Due to the likelihood of women to interact with other women, they suggest gender-aware recruiting to address the gender gap.

The authors point out the utility of positive emotion in keeping discussions on track, and suggest that experienced editors should be encouraged to maintain a positive climate. To determine users' gender, they used a crowd-sourced study through Crowdflower. Emotions are determined using the ANEW wordlist which distinguishes the range of emotional variability, based on valence, arousal, and dominance. The paper notes that policy mentions tend to have "a remarkably positive and dominant tone, and with stronger emotional load than in the rest of the discussion'".

Editor collaboration patterns

A paper from the University of Alberta addresses the difficulty with analyzing edit histories and finding conflict in particular [8]. They use terms indicating content-based agreement (e.g. "add", "fix", "spellcheck", "copy", and "move") and disagreement ("uncited", "fact", "is not", "bias", "claim", "revert", and "see talk page"). They define conflicting interactions as those that revert, or delete content, or use more negative terms than positive terms. They find that this is a useful way to identify controversial articles.

Why does the number of Wikipedia readers rise while the number of editors doesn't?

A student paper for a course on "Project in Mining Massive Data Sets [9]" at Stanford University, titled "Wikipedia Mathematical Models and Reversion Prediction"[10] tries to use mathematical models "to explain why the amount of [editors on the English Wikipedia] stops increasing, whereas the amount of viewers keeps increase", and "to predict if an edit will be reverted." The researchers used Elastic MapReduce on Amazon's servers to carry out this research. The paper is a bit confused since the researchers are more interested in models and validation than explaining the phenomena.

The first part of the paper includes two models for examining the relation of visitors to editors in Wikipedia's community. The first model makes the assumption that editors act as predators and articles have the role of prey. However this model did not fit the data. The second model used a linear regression between a number of factors which allow to model the community's statistics over time. The model is then tested using simulation and seems to present accurate results.

In the second part of the paper, three models were used predict which edits will get reverted. The models were trained using 24 features, classified either as edit, editor or article based. E.g. an article's age; its edit count; number of editors participating in editing; number of articles the editor has edited; change in information compared to previous status. The outcome of the prediction which used three machine learning algorithms achieved about 75% accuracy and another interesting conclusion was that the ability to detect reversion has not changed much over time.

Briefly

- **What was the most influential paper ever about Wikipedia and related topics?**: Wikimedia France is currently seeking nominations for its Research Award (which comes with a grant of €2500), which aims "to reward the most influential research paper on Wikimedia projects" published between 2003 and 2011. In the coming years, the scope is to be widened to include free knowledge projects more generally. Submission deadline for paper nominations is **August 7**. The winner shall be announced in November.
- **Retrieving information missing from Wikipedia articles**: A paper presented at the *6th International Conference on Ubiquitous Information Management and Communication* presents a technique developed by researchers at Kyoto University to compare Wikipedia articles with matching sources retrieved via search engines and identify, via topic modeling, to what extent the external source includes complementary information not covered in the article.[11] The paper then proposes a method to extract sentences from these sources and rank them to facilitate editorial work. Two case studies are discussed analyzing the *Yutaka Taniyama* and *Influvac* articles from the English Wikipedia.
- **Mining Wikipedia for common traits of notable individuals**: Researcher Pauline C. Ng presented a paper at *ICWSM '12* showcasing the potential of using Wikipedia as a corpus of data to study the common characteristics of "notable individuals".[12] Names and birth locations of a list of 40,250 people born in the United States from 1940–1989 and with a Wikipedia article were compared against census data. The analysis reveals interesting patterns such as the fact that "people with rare names [are] more than 2x likely to appear in Wikipedia" or that "people with nicknames are more likely to be in Wikipedia", but with a significantly more pronounced effect for male than female individuals. The author suggests that mining Wikipedia biographies may help "discover novel characteristics associated with positive life outcomes". The main findings of the paper are summarized in this blog post [13].
- **2012 Aurora shooting**: Brian Keegan, who has published a series of previous articles on coverage of breaking news topics in Wikipedia (see e.g. our past coverage: "High-tempo contributions: Who edits breaking news articles?"), published a series an analyses and a series of graphs on the first several days of responses and article writing on Wikipedia to cover the 2012 Aurora shootings on English Wikipedia.[14] Several participants responded to Keegan in comments on his blog. Taha Yasseri published a graph of the increase in the number of articles on the shootings in different languages.[15]
- **Detecting featured articles using fuzzy logic**: A paper[16] by two Bangkok-based computer scientists constructed a fuzzy logic ruleset to discern the featured articles on the Thai Wikipedia (88 at the time of the study) from non-featured articles (100 in the examined sample). Using 26 different rules, from unsurprising ones such as the assumption that an article with few footnotes probably does not have featured status, to more complicated criteria involving the most frequent and second most frequent editor of the article, they achieved 100% recall (i.e. detecting all featured articles) and 86% precision (i.e. of the articles detected as having featured quality, 86% actually had featured article status). This compared favorably to a different detection method (which clustered articles according to their distance in a similarity measure that the authors do not specify), supporting the authors' thesis that fuzzy logic is a better approach to the problem, because "the quality of Wikipedia articles should be graded [by] more than two values (good or not good)". (See also coverage of an earlier paper with similar goal: "Lexical clues" predict article quality)

References

[1] Török, J.; Iñiguez, G.; Yasseri, T.; San Miguel, M.; Kaski, K.; Kertész, J. (2012) " Opinions, Conflicts and Consensus: Modeling Social Dynamics in a Collaborative Environment (http://arxiv.org/abs/1207.4914v1)". ArXiv. Open access

[2] Nagaraj, Abhishek. (2012) " The effect of copyright law on the reuse of digital content (https://wikimania2012.wikimedia.org/wiki/File:Wikimania_nagaraj.pdf)". *Wikimania 2012*, July 12–15 2012, George Washington University. Open access

[3] http://www.theatlantic.com/technology/archive/12/07/exactly-how-copyright-laws-impoverish-some-of-wikipedia-as-calculated-by-an-mit-economist/259970/

[4] (2012) "From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks". *ZooKeys* 209: 235. : 10.3897/zookeys.209.3247 (http://dx.doi.org/10.3897/zookeys.209.3247). Open access

[5] http://ipt.vertnet.org:8080/ipt/resource.do?r=hendersonnotebooks1-3

[6] (2012) "Quality of Internet information in pediatric otolaryngology: A comparison of three most referenced websites". *International Journal of Pediatric Otorhinolaryngology*. : 10.1016/j.ijporl.2012.05.026 (http://dx.doi.org/10.1016/j.ijporl.2012.05.026). Closed access

[7] Laniado, David; Castillo, Carlos; Kaltenbrunner, Andreas; Fuster Morell, Mayo (Aug 27–29, 2012). "Emotions and dialogue in a peer-production community: the case of Wikipedia" (http://chato.cl/papers/laniado_kaltenbrunner_castillo_fuster_2012_emotions_wikipedia.pdf). *WikiSym*. Linz, Austria: ACM Press. .Open access

[8] Sepehri-Rad, Hoda; Makazhanov, Aibek; Rafiei, Davood; Barbosa, Denilson. (2012) " Leveraging Editor Collaboration Patterns in Wikipedia (http://webdocs.cs.ualberta.ca/~drafiei/papers/ht12.pdf)". Open access In Proceedings of the 23rd ACM conference on Hypertext and Social Media, pp. 13-22. doi:10.1145/2309996.2310001 Closed access

[9] http://www.stanford.edu/class/cs341/reports/index.html

[10] Jia Ji; Bing Han; Dingyi Li. (2012) " Wikipedia Mathematical Models and Reversion Prediction (http://www.stanford.edu/class/cs341/reports/15-Jia_FinalReport_Team15.pdf)" Open access

[11] Eklou, D., Asano, Y., & Yoshikawa, M. (2012). How the web can help Wikipedia: a study on information complementation of Wikipedia by the web. *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication – ICUIMC '12* (p. 1). New York, New York, USA: ACM Press. doi:10.1145/2184751.2184763 Closed access

[12] Ng, P. C. (2012). " What Kobe Bryant and Britney Spears Have in Common: Mining Wikipedia for Characteristics of Notable Individuals (https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4556)". *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Open access

[13] http://paulinepi.com/2012/06/mining-wikipedia-paper-at-icwsm-2012/

[14] Keegan, Brian. (July 21, 2012) " Aurora shootings (http://www.brianckeegan.com/2012/07/2012-aurora-shootings/2012)."

[15] Yasseri, Taha. (2012) " Number of covering WPs vs. time (http://wwm.phy.bme.hu/blog.html)" (http://wwm.phy.bme.hu/figs/aurora.png).

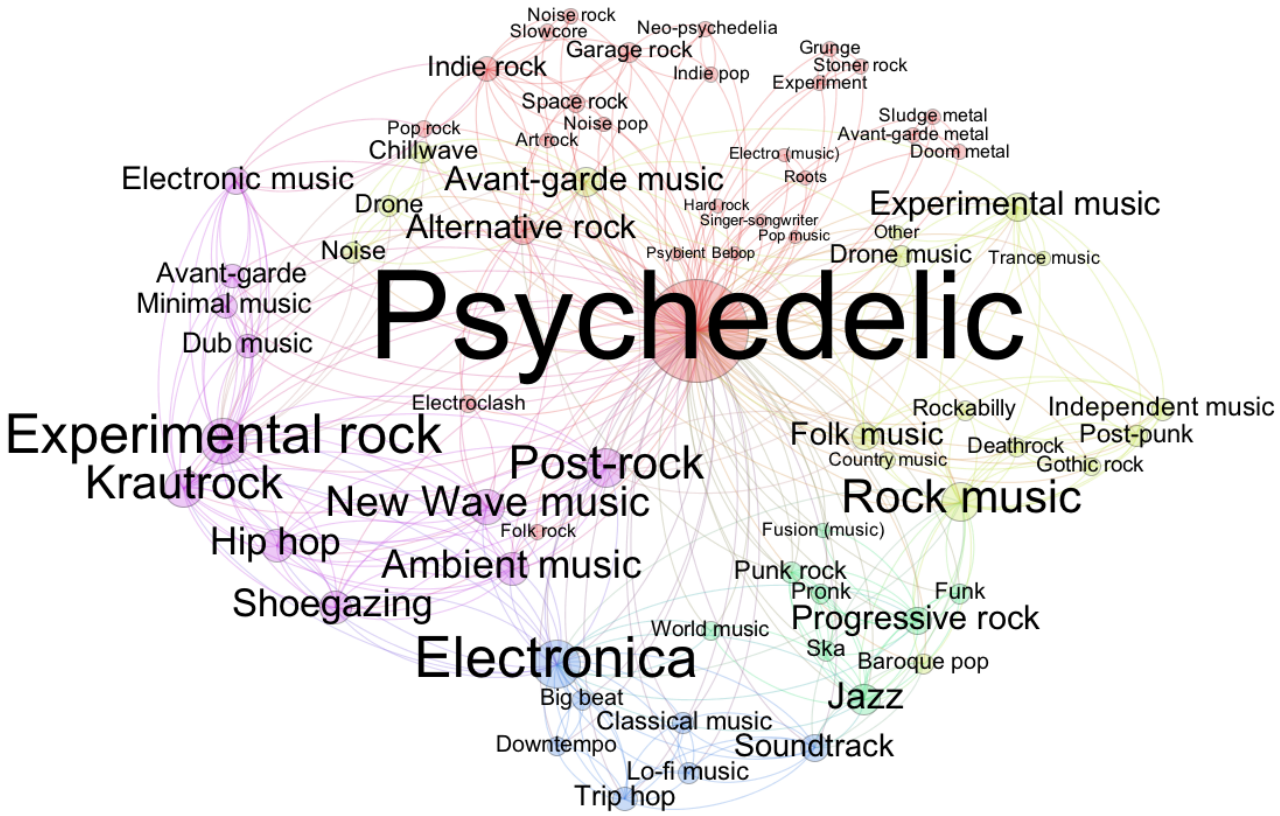
[16] Saengthongpattana, Kanchana; Soonthornphisaj, Nuanwan. (2012) " Thai Wikipedia Quality Measurement using Fuzzy Logic (https://kaigi.org/jsai/webprogram/2012/pdf/697.pdf)" *26th Annual Conference of the Japanese Society for Artificial Intelligence*, June 12-15, 2012, Yamaguchi, Japan. Open access

Issue 2(8): August 2012

New influence graph visualizations; NPOV and history; 'low-hanging fruit'

With contributions by: Piotrus, Ragesoss, Evan, DarTar, Tbayer and OrenBochman

Wikipedia-based graphs visualize influences between thinkers, writers and musicians



A visualization of musical genres related to psychedelic music [1], based on DBPedia data.

In a blog post titled "Graphing the history of philosophy",[2] Simon Raper of the company MindShare UK describes how he constructed an influence graph of all philosophers using the "Influenced by" and "Influenced" fields of Template:Infobox philosopher (example: Plato). This information was retrieved using DBpedia with a simple SPARQL query. After some cleanup, the result, consisting of triplets in the form <Philosopher A, Philosopher B, Weight> was processed using the open source graph visualization package Gephi to create an impressive overview of the philosophers within their respective spheres of influence.

Brendan Griffen extended the idea to "*everyone* on Wikipedia. Well, everyone with an infobox containing 'influences' and/or 'influenced by'", arriving at a huge, far more dense "Graph Of Ideas" including not only philosophers, but also novelists, fantasy and science fiction writers, and comedians.[3] In another blog post,[4] Griffen added transitive links as well – so that each person is considered to be influenced both directly and indirectly. The most connected people in the graph were ancient Greek thinkers, with Thales, Pythagoras and Zeno of Elea occupying the top three spots. Griffen remarks that this vindicates a statement in Bertrand Russell's History of Western Philosophy (1945): "Western Philosophy begins With Thales".

Also inspired by Raper's posting, Tony Hirst posted a number of visualizations of the Wikipedia link and category structure (likewise using DBpedia and Gephi, queried via the Semantic Web Import plugin [5]) to visualize related entries and influence graphs in the English Wikipedia. The blog posts (all of which include detailed step-by-step tutorials) examine the related graph of philosophers,[6] and also visualize an influence graph of programming

languages[7] and one of musical genres related to psychedelic music.[8] All these visualizations and blog posts by Hirst are released under a Creative Commons Attribution license [9].

Hirst also mentioned a related tool called "WikiMaps", the subject of a recent article in the International Journal of Organisational Design and Engineering.[10] As described in a press release [11], the tool provides a "map of what is “important” on Wikipedia and the connections between different entries. The tool, which is currently in the “alpha” phase of development, displays classic musicians, bands, people born in the 1980s, and selected celebrities, including Lady Gaga, Barack Obama, and Justin Bieber. A slider control, or play button, lets you move through time to see how a particular topic or group has evolved over the last 3 or 4 years." A demo version [12] is available online.

See also the recent coverage of a similar visualization, based on wikilinks instead of infoboxes: "The history of art mapped using Wikipedia"

Information retrieval scientists turn their attention to Wikipedia's page view logs

The Time-aware Information Access workshop [13] at this year's SIGIR (Special Interest Group on Information Retrieval) conference brought a wave of attention to Wikipedia's public page-view logs. Detailing the number of page views per hour for every Wikipedia project, these files [14] figure prominently in a variety of open-source intelligence applications presented at the workshop.

Found to be connected to the "#euro2012" hashtag by analyzing Wikipedia pageviews: Euro 2012 football championship

A group of researchers from ISLA, University of Amsterdam created an API providing access to this data and performing simple analysis tasks.[15] Though the site [16] appears to be down at the time of writing, the API supports the retrieving a particular article's page-view time series as well as searching for other wikipedia articles based on the similarity of their time series. In addition to machine-readable JSON results, the API will supply simple plots in png format. While the idea of providing page specific time series is not new, support for finding other pages with similar viewing patterns highlights a fascinating new use for Wikipedia page views.

Two other papers are combining Wikipedia page-view information with external time-series data sets. On the intuition that Wikipedia page views should have a strong correlation with real-world events, researchers from the University of Glasgow and Microsoft built a system to detect which hashtags frequently queried on Bing Social Search were event-related.[17] For example, the hashtag #thingsthatannoyme doesn't clearly correspond to an event, whereas a hashtag like "#euro2012" is about the UEFA European Football Championship. After tokenizing the hashtags into a list of words, the researchers queried Wikipedia for those terms and correlated the time series of hashtag search popularity with the page-view time series for the articles which are returned. This correlation score can be used to indicate which hashtags are likely to be about events, a useful feature for web searches and any other temporally aware zeitgeist application.

In a similar vein, researchers from the University of Edinburgh and University of Glasgow used the Wikipedia page-view stream to tackle the problem known as first-story detection (FSD), which aims to automatically pick out the first publication relating to a new topic of interest.[18] While traditional techniques primarily focus on newswire or Twitter, the authors used a combination of Twitter and Wikipedia page views to construct an improved FSD system. To improve on state-of-the-art Twitter-only FSD systems, the authors aimed to filter out false positives by checking that the Twitter-based first stories corresponded to a Wikipedia page that was also experiencing heightened traffic during the same period.

Using a simple outlier detection method, the authors created a set of Wikipedia pages with unexpectedly high page views for each hour. Each Twitter-based first story (tweet) was then matched against the corresponding collection of Wikipedia outliers, employing an undisclosed metric of textual similarity that uses only the Wikipedia page titles. If the tweet failed to match any spiking Wikipedia page, it was down-weighted as a first story candidate. The authors showed that this combined approach improves FSD precision in comparison to a twitter-only baseline for all but the most popular twitter-based stories. Though this research makes advances on the difficult task of first-story detection, perhaps the most immediately useful finding is that Wikipedia page views appear to lag behind twitter activity by roughly two hours. In general, we can expect to see an increasing amount of joint models over various open-source intelligence streams as we learn exactly what each stream is useful for and the relationships between the streams.

See also the *Signpost* coverage of a small study of the highest hourly page views on the English Wikipedia during January-July 2010, and their likely causes: "Page view spikes"

The limits of amateur NPOV history

In "The inclusivity of Wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists"[19], information studies professor Brendan Luyt of Nanyang Technological University looks at History of the Philippines, a B-class article that had featured article status from October 2006 until it was delisted at the conclusion of its featured article review in January 2011.

Luyt argues that talk-page discussions, the types of sources cited, and the organization of the article itself, all point to a very traditional view of what constitutes history: in short, great man history concerned mainly with political and military events, and the actions of elites. This style of history does not capture the breadth of approaches used by professional historians, so does not live up to the ideal of NPOV in which all significant viewpoints published in reliable sources are represented fairly and proportionately. In practice, Luyt shows, editors (lacking sufficient knowledge of the relevant professional historical literature) end up using arguments over bias and NPOV to construct a limited and conservative historical narrative—for this article at the least, although a similar pattern could be found for many broad historical topics.

The sources cited are primarily what Luyt calls "textbookese" summaries, easily available online, which focus on bare facts without the historical debates that surround them. Between the valid sources and experts recognized by Wikipedia editors and the good-faith use of the NPOV principle to limit other viewpoints, Luyt concludes that—rather than being more inclusive of diverse views and sources than the typical "expert" community—Wikipedia in practice recognizes a considerably narrower set of viewpoints.

Three new papers about Wikipedia class assignments

An article titled "Assigning Students to edit Wikipedia: Four Case Studies"[20] presents the experiences of four professors who participated in the Wikipedia Education Program, in a total of six courses total (two of four instructors taught two classes each). The lessons from the assignments included: 1) the importance of strict deadlines, even for graduate classes; 2) having a dedicated class for acquiring skills in editing and for understanding Wikipedia policies, or spreading this over segments of several classes; 3) the benefits of having students interact with the campus ambassadors and the wider Wikipedia community.

Overall, the instructors saw that compared with their engagement in traditional assignments, students were more highly motivated, produced work of higher quality, and learned more skills (primarily, related to using Wikipedia, such as being able to better judge its reliability). Wikipedia itself benefited from several dozen created or improved articles, a number of which were featured as DYKs. The paper presents a useful addition to the emerging literature on teaching with Wikipedia, as one of the first serious and detailed discussions of specific cases of this new educational approach.

"Integrating Wikipedia Projects into IT Courses: Does Wikipedia Improve Learning Outcomes?"[21] is another paper that discusses the experiences of instructors and students involved in the recent Wikipedia:Global Education

Program. Like most existing research in this area, the paper is roughly positive in its description of this new educational approach, stressing the importance of deadlines, small introductory assignments familiarizing students with Wikipedia early in the course, and the importance of close interactions with the community. A poorly justified (or explained) deletion or removal of content can be quite a stressful experience to students (and the newbie editors are unlikely to realize that an explanation may be left in an edit summary or page-deletion log). A valuable suggestion in the paper was that instructors (professors) make edits themselves, so they would be able to discuss editing Wikipedia with students with first-hand experience instead of directing students to ambassadors and how-to manuals; and to dedicate some class time to discussing Wikipedia, the assignment, and collective editing.

A four-page letter[22] in the *Journal of Biological Rhythms* by a team of 48 authors reported on a a similar undergraduate class project [23] in early 2011, where 46 students edited 15 Wikipedia articles in the field of chronobiology, aiming at good article status. After their first edits, they were systematically given feedback by one "Wikipedia editor and 6 experts in chronobiology" before continuing their edits (in the paper's acknowledgements the authors also thank "innumerable Wikipedia editors who critiqued student edits"). Because of the high visibility of the results – most of the articles were ranked top in Google results – students found the experience rewarding. Topics were selected collaboratively by the class, and because students came up with a relatively small number of suggestions, one concern was that the project might, if repeated, run out of article topics in the given subject area.

A literature review presented at July's Worldcomp'12 [24] conference in Las Vegas about "Wikipedia: How Instructors Can Use This Technology As A Tool In The Classroom"[25] also recommended to have students actively edit Wikipedia (as well as practicing to read it critically), and concluded that "it is time to embrace Wikipedia as an important information provider and one of the innovative learning tools in the educators' toolbox."

Substantive and non-substantive contributors show different motivation and expertise

"Investigating the determinants of contribution value in Wikipedia"[26] reports the results of a survey of Wikipedians who were asked their opinion about the "contribution value" of their edits (measured by agreement to statements such as "your contribution to Wikipedia is useful to others"), which was then related to various characteristics.

The researchers used Google to obtain a list of 1976 Wikipedia users' email addresses (using keywords such as "gmail.com" or "hotmail.com"). They sent invitation emails that provided the URL to the online questionnaire. In six weeks, 234 editors completed all the questions. Of these, 205 – Nine females and 196 males – supplied a valid user name and were considered in the rest of the analysis (anonymous editors were removed).

A content analysis was performed of 50 randomly selected edits by each respondent (or all, if the user had fewer than 50 edits), classifying them as "substantive" changes (e.g. "add links, images, or delete inaccurate content") and "non-substantive changes" (e.g. "reorganizing existing content [or] correcting grammatical mistakes and formatting texts to improve the presentation"), corresponding to "two [proposed] new contributor types in Wikipedia to discriminate their editing patterns."

An attempt was made to relate this to the "contribution value" the respondents assigned to their own edits, and to their responses in two other areas:

- "interests" (measured by respondent ratings of a variety of different motivations to contribute to Wikipedia on how well each applied to themselves, e.g. "Enhancing your learning abilities, skills and expertise"); and
- "resources" (meaning expertise based on education, profession and hobbies, measured by respondent ratings of their expertise in a variety of fields within eachc, e.g. "Hospitality and tourism").

The "breadth" of interests and resources was defined as the number of ratings above a certain threshold in each, and the "depth" as the highest rating assigned in each.

In an "important consideration for practitioners", the authors wrote that:

> *"[T]o produce valuable contributions, users with high depth of interests and resources should be encouraged to concentrate their efforts on substantive changes. Meanwhile, for users with high breadth of interests and*

> *resources, wiki practitioners should advise them to pay more attention to nonsubstantive changes. The findings imply that practitioners can try to identify two distinct types of users. To achieve this objective, they may develop certain algorithms in wikis to automatically detect the frequencies of substantive/non-substantive changes of users. ... For example, notification messages about wiki articles that need substantive changes can be sent to users who have high levels of depth of interests and resources. Similarly, well-prepared messages about articles that need non-substantive changes can be delivered to users who have high levels of breadth of interests and resources."*

Is there systemic bias in Wikipedia's coverage of the Tiananmen protests?

Wikipedia: Remembering in the digital age[27] is a masters dissertation by Simin Michelle Chen, examining collective memories as represented on the English Wikipedia; she looked at how significant events are portrayed (remembered) on the project, focusing on the Tiananmen Square Protests of 1989. She compared how this event was framed by the articles by New York Times and Xinhua News Agency, and in Wikipedia, where she focused on the content analysis of w:Talk:Tiananmen Square protests of 1989 and its archives.

Remembrance of the 20th anniversary of the June 4 events in Hong Kong (replica of the "goddess of democracy" statue)

Chen found that the way Wikipedia frames the event is much closer to that of *The New York Times* than the sources preferred by the Chinese government, which, she notes, were "not given an equal voice" (p. 152). This English Wikipedia article, she says, is of major importance to China, but is not easily influenced by Chinese people, due to language barriers, and discrimination against Chinese sources that are perceived by the English Wikipedia as unreliable – that is, more subject to censorship and other forms of government manipulation than Western sources. She notes that this leads to on-wiki conflicts between contributors with different points of views (she refers to them as "memories" through her work), and usually the contributors who support that Chinese government POV are "silenced" (p. 152). This leads her to conclude that different memories (POVs) are weighted differently on Wikipedia. While this finding is not revolutionary, her case study up to this point is a valuable contribution to the discussion of Wikipedia biases.

While Chen makes interesting points about the existence of different national biases, which impact editors very frames of reference, and different treatment of various sources, her subsequent critique of Wikipedia's NPOV policy is likely to raise some eyebrows (pp. 48–50). She argues that NPOV is flawed because "it is based on the assumption that facts are irrefutable" (p. 154), but that those facts are based on different memories and cultural viewpoints, and thus should be treated equally, instead of some (Western) being given preference. Subsequently, she concludes that Wikipedia contributes to "the broader structures of dominance and Western hegemony in the production of knowledge" (p. 161).

While she acknowledges that official Chinese sources may be biased and censored, she does not discuss this in much detail, and instead seems to argue that the biases affecting those sources are comparable to the those affecting Western sources. In other words, she is saying that while some claim Chinese sources are biased, other claim that Western sources are biased, and because the English Wikipedia is dominated by the Western editors, their bias triumphs – whereas ideally, all sources should be acknowledged, to reduce the bias. The suggestion is that Wikipedia should reject NPOV and accept sources currently deemed as unreliable. Her argument about the English Wikipedia having a Western bias is not controversial, was discussed by the community before (although Chen does not seem to be aware of it, and does not use the term "systemic bias" in her thesis) and reducing this bias (by improving our coverage of non-Western topics) is even a goal of the Wikimedia Foundation. However, while she does not say so directly, it appears to this reviewer that her argument is: "if there are no reliable non-Western sources, we should use the unreliable ones, as this is the only way to reduce the Western bias affecting non-Western topics". Her ending comment that Wikipedia fails to leave to its potential and to deliver "postmodern approach to truth" brings to mind the community discussions about verifiability not truth (the existence of this debates she briefly acknowledges on p. 48).

Remembrance in the West (replica of the same statue at the University of British Columbia, Canada)

Overall, Chen's discussion of biases affecting Wikipedia in general, and of Tiananmen Square Protests in particular, is useful. The thesis however suffers from two major flaws. First, the discussion of Wikipedia's policies such as reliable sources and verifiability (not truth ...) seems too short, considering that their critique forms a major part of her conclusions. Second, the argumentation and accompanying value-judgements that Wikipedia should stop discriminating against certain memories (POVs) is not convincing, lacking a proper explanation of the reasons why the Wikipedia community made those decisions favoring verifiability and reliable sources over inclusion of all viewpoints. Chen argues that Wikipedia sacrifices freedom and discriminates against some memories (contributors), which she seems to see as more of a problem that if Wikipedia was to accept unreliable sources and unverifiable claims.

"Low-hanging fruit hypothesis" explains Wikipedia's slowed growth?

A student paper titled "Wikipedia: nowhere to grow"[28] from a Stanford class about "Mining Massive Data Sets" [29] argues for the "low-hanging fruit hypothesis" as one factor explaining the well-known observation that "since 2007, the growth of English Wikipedia has slowed, with fewer new editors joining, and fewer new articles created". The hypothesis is described as follows: "the larger [Wikipedia] becomes, and the more knowledge it contains, the more difficult it becomes for editors to make novel, lasting contributions. That is, all of the easy articles have already been created, leaving only more difficult topics to write about". The authors break this hypothesis into three smaller ones that are easier to test – that (1) there has been a slowing in edits across many languages with diverse characteristics; (2) older articles are more popular to edit; and (3) older articles are more popular to read. They find a support for all three of the smaller hypotheses, which they argue supports their main low-hanging fruit hypothesis.

While the overall study seems well-designed, the extrapolation from the three subhypotheses to the parent hypothesis seems problematic. The authors do not provide a proper operationalization of terms such as "novel", "lasting", and "easy/difficult", making it difficult to enter into a discourse without risking miscommunication. There may be at least four main issues in the work:

- (a minor but annoying issue): hypothesis II is incorrectly and confusingly worded in the section dedicated to it: "Older articles (those created earlier) will be more popular to read than more newly created articles"; however, their study of hypothesis II is based on the number of edits to the article, not the number of page views (those are analyzed in the subsequent hypothesis III);
- regarding the claim "all of the easy articles have already been created, leaving only more difficult topics to write about", it is true that the majority of vital/core articles are developed beyond stub, and their subsequent expansion is more difficult (it takes more and more effort to move the article up through assessment classes). However, while the older articles are more popular, they are not necessarily easier to edit, as w:Wikipedia:The Core Contest illustrates. While almost everyone may be able to quickly define (stub) Albert Einstein, it is questionable whether 1) developing this article is easier than developing an article on a less well-known subject, where fewer sources mean the editors need to do less research, and 2) while mostly everyone knows who Einstein was, everyone also has knowledge of at least *some* less popular subjects. As Wikipedia:Missing articles illustrate, there are still many articles in need of creation, and for a fan/expert, it may be easier to create an article on an esoteric subject than to edit the article on Einstein.
- The claim that "[it is more difficult] for editors to make novel, lasting contributions" is difficult to analyze due to the lack of operationalization of those terms by the authors, but 1) regarding novel, if it means new, see the Missing articles argument above – there is still plenty to write about; and 2) regarding lasting – the authors do not cite any sources suggesting the deletionism in English Wikipedia may be on the rise.

Overall, the paper presents four hypotheses, three of which seem to be well supported by data, and contribute to our understanding of Wikipedia, but their main claim seems rather controversial and poorly supported by their data and argumentation.

See also the coverage of a related paper in a precursor of this research report last year: "IEEE magazine summarizes research on sustainability and low-hanging fruit"

Briefly

- **Barnstars at ASA annual conference**: Two Wikipedia papers were presented at the 2012 annual meeting of the American Sociological Association last week, both focusing on "barnstar" awards on Wikipedia.
  Michael Restivo and Arnout van de Rijt presented their research on the effect of barnstars, titled "Experimental Study of Informal Rewards in Peer Production", which had found that assigning "editing awards or 'barnstars' to a subset of the 1% most productive Wikipedia contributors ... increases productivity by 60% and makes contributors six times more likely to receive additional barnstars from other community members", as stated in the abstract. See the review in the April issue of this report: "Recognition may sustain user participation".
  Benjamin Mako Hill, Aaron Shaw, and Yochai Benkler presented "Status, Social Signaling, and Collective Action: A Field Study of Awards on Wikipedia", with a more skeptical look at the effect of barnstars. According to the abstract, "Willer has argued for a sociological mechanism for the provision of public goods through selective incentives. Willer posits a "virtuous circle" in which contributors are rewarded with status by other group members and in response are motivated to contribute more. [... But] there is reason to suspect that not all individuals will be equally susceptible to status-based awards or incentives. At the very least, Willer's theory fails to take into account individual differences in the desire to signal contributions to a public good. We test whether this omission is justified and whether individuals who do not signal status in the context of collective action behave differently from those who do in the presence of a reputation-based award. [Analyzing barnstars on Wikipedia,] we show that the social signalers see a boost in their editing behavior where non-signalers do not."

- **How high school, college and PhD students evaluate Wikipedia quality**: "Trust in online information A comparison among high school students, college students and PhD students with regard to trust in Wikipedia"[30] is a master thesis that looks at how these three groups judge the trustworthiness of Wikipedia articles, based on the "3S-model" model by the advisors of the thesis (Lucassen and Schraagen (2011), *Factual Accuracy and Trust in Information: The Role of Expertise. Journal of the American Society for Information Science & Technology,* 62, 1232–1242). Unsurprisingly, the more educated the group is, the more detailed their analysis will be. High school students usually focus on accuracy, completeness, images, length, and writing style. College and PhD students go beyond those five elements, although looking at authority, objectivity, and structure. Interestingly, the differences between college and PhD students were much smaller than those between high school students and the other two groups. Another important finding of the study was that the less educated the group, the less likely they are to be aware of Wikipedia being open source and open to editing by anyone. Further, high school students seem to have much more difficulty in distinguishing between a high and low quality article, and overall, seem much more likely to simply not question the trustworthiness of the sources given.
- **Doctors widely use Wikipedia as a reference**: A literature review of 50 articles about the use of social media by clinicians[31] found that "Wikipedia is widely used as a reference tool" among them, despite concerns about its accuracy. The authors remark that "we found multiple projects that sought to emulate Wikipedia's success in crowd-sourcing useful medical content, while additionally emphasizing editorial credibility by verifying credentials of contributors. These include RadiologyWiki, announced in 2007 and currently dormant, and Medpedia, which launched in 2009 with substantial institutional backing. We did not find articles reporting success metrics for these projects or similar ones."
- **Predicting quality flaws in Wikipedia articles**: A notebook paper to presented at the annual PAN workshop at the *Conference and Labs of the Evaluation Forum* meeting (CLEF '12) introduces *FlawFinder*, a toolset to predict quality flaws in Wikipedia articles.[32] The paper is one of the winning entries in a Competition on Quality Flaw Prediction in Wikipedia [33]. The paper defines 11 types of quality flaws, spanning low-level issues (such as *orphaned* or *unreferenced articles*) and high-level quality flaws (such as *notability* or *original research*). It uses a corpus of articles tagged with cleanup templates (154,116 articles from a January 2012 dump of the English Wikipedia) as a training set to predict whether articles in a separate, uncategorized set suffer from the same flaws. The model uses a variety of features of the training set based on revision data, lexical properties, structural properties of the article and the reference section, network properties of the link graph. The results suggest, among other things, that the strongest non-lexical features for the *advert* flaw are links pointing to external resources, while the number of discussions on article's talk page is the strongest feature to predict *original research*.
- **Quality of text and quality of editors**. A poster presented at the *2012 ACM Conference on Hypertext and Social Media* (HT 2012) describes a method to measure the quality of Wikipedia articles by combining text survival metrics and the quality of editors editing these articles, where editor quality is calculated recursively as a function of the quality of their contributions. The method claims to be "resistant to vandalism", however no empirical validation is presented in the poster.[34]
- **WikiSym 2012**: WikiSym, the annual conference "dedicated to wiki and open collaboration research and practice" was happening in Linz, Austria as this issue of the research report went to press. Links to online versions of all conference papers have been posted in the program [35]; expect fuller coverage in the September issue.

References

[1] http://blog.ouseful.info/2012/07/04/mapping-related-musical-genres-on-wikipediadbpedia-with-gephi/

[2] Raper, Simon: Graphing the history of philosophy (http://drunks-and-lampposts.com/2012/06/13/graphing-the-history-of-philosophy/). Drunks and Lampposts, June 13, 2012

[3] Brendan Griffen: The Graph Of Ideas (http://griffsgraphs.com/2012/07/03/graphing-every-idea-in-history/). Griff's Graphs, July 3, 2012

[4] Brendan Griffen: The Graph Of Ideas 2.0 (http://griffsgraphs.com/2012/07/20/the-graph-of-ideas-2-0/). Griff's Graphs, July 20, 2012

[5] http://wiki.gephi.org/index.php/SemanticWebImport

[6] Hirst, Tony (2012). Visualising related entries in Wikipedia using Gephi (http://blog.ouseful.info/2012/07/03/visualising-related-entries-in-wikipedia-using-gephi/). OUseful.Info, July 3, 2012

[7] Hirst, Tony (2012). Mapping how Programming Languages Influenced each other According to Wikipedia (http://blog.ouseful.info/2012/07/03/mapping-how-programming-languages-influenced-each-other-according-to-wikipedia/). OUseful.Info, July 3, 2012

[8] Hirst, Tony (2012). Mapping related Musical Genres on Wikipedia with Gephi (http://blog.ouseful.info/2012/07/04/mapping-related-musical-genres-on-wikipediadbpedia-with-gephi/). OUseful.Info, July 4, 2012

[9] http://creativecommons.org/licenses/by/3.0/

[10] "Wikimaps: dynamic maps of knowledge" in Int. J. Organisational Design and Engineering, 2012, 2, 204–224

[11] http://sciencespot.co.uk/mapping-research-on-wikipedia-with-wikimaps.html

[12] http://www.ickn.org/wikimaps/

[13] http://research.microsoft.com/en-us/people/milads/taia2012.aspx

[14] http://dumps.wikimedia.org/other/pagecounts-raw/

[15] Peetz, M. H., Meij, E., & de Rijke, M. (2012). OpenGeist: Insight in the Stream of Page Views on Wikipedia. *SIGIR 2012 Workshop on Time-aware Information Access* (#TAIA2012). **PDF** (https://research.microsoft.com/en-us/people/milads/taia2012-opengeist-self-contained.pdf) Open access

[16] http://www.opengeist.org/

[17] Whiting, S., Alonso, O., & View, M. (2012). Hashtags as Milestones in Time. *SIGIR 2012 Workshop on Time-aware Information Access* (#TAIA2012). **PDF** (https://research.microsoft.com/en-us/people/milads/hashtagsasmilestonesintime.pdf) Open access

[18] Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., & Ounis, I. (2012). Bieber no more: First Story Detection using Twitter and Wikipedia (http://www.dcs.gla.ac.uk/~craigm/publications/osborneTAIA2012.pdf). SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012). Open access

[19] Luyt, B. (2012). The inclusivity of Wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists. Journal of the American Society for Information Science and Technology, 63(9), 1868–1878. doi:10.1002/asi.22671 Closed access

[20] Carver, B., Davis, R., Kelley, R. T., Obar, J. A., & Davis, L. L. (2012). Assigning Students to Edit Wikipedia: four case studies. E-Learning and Digital Media, 9(3), 273–283. PDF (http://www.wwwords.co.uk/rss/abstract.asp?j=elea&aid=5094) Closed access

[21] Patten, K., & Keane, L. (2012). Integrating Wikipedia Projects into IT Courses: Does Wikipedia Improve Learning Outcomes? AMCIS 2012 Proceedings. PDF (http://aisel.aisnet.org/amcis2012/proceedings/EndUserIS/28) Open access

[22] Chiang, C. D., Lewis, C. L., Wright, M. D. E., Agapova, S., Akers, B., Azad, T. D., Banerjee, K., et al. (2012). Learning chronobiology by improving Wikipedia. *Journal of Biological Rhythms*, 27(4), 333–36. **HTML** (http://www.ncbi.nlm.nih.gov/pubmed/22855578) Closed access

[23] http://www.nslc.wustl.edu/courses/Bio4030/wikipedia_project.html

[24] http://www.world-academy-of-science.org/worldcomp12/ws/program/sww17

[25] Hogg, J. L. (2012). Wikipedia: How Instructors Can Use This Technology As A Tool In The Classroom. Worldcomp'12. **PDF** (http://elrond.informatik.tu-freiberg.de/papers/WorldComp2012/EEE6068.pdf) Open access

[26] Zhao, S. J., Zhang, K.Z.K., Wagner, C., & Chen, H. (2012). Investigating the determinants of contribution value in Wikipedia. *International Journal of Information Management*. doi:10.1016/j.ijinfomgt.2012.07.006 Closed access

[27] Chen, Simin Michelle (2012): Wikipedia: Remembering in the digital age. University of Minnesota MA thesis. June 2012. **PDF** (http://purl.umn.edu/131343) Open access

[28] Austin Gibbons, David Vetrano, Susan Biancani (2012). Wikipedia: Nowhere to grow (http://www.stanford.edu/class/cs341/reports/09-GibbonsVetranoBiancaniCS341.pdf) Open access

[29] http://www.stanford.edu/class/cs341/

[30] Rienco Muilwijk: Trust in online information A comparison among high school students, college students and PhD students with regard to trust in Wikipedia. University of Twente, February 2012 PDF (http://essay.utwente.nl/61631/1/Muilwijk,_M.C._-_s0150908_(verslag).pdf) Open access

[31] von Muhlen, M., & Ohno-Machado, L. (2012). Reviewing social media use by clinicians. Journal of the American Medical Informatics Association : JAMIA, 19(5), 777–81. doi:10.1136/amiajnl-2012-000990 Open access

[32] Ferschke, O., Gurevych, I., & Rittberger, M. (2012). FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. *Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) Workshop* (PAN @CLEF 2012), Rome. **PDF** (http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/Ferschke2012_FlawFinder_CLEF.pdf) Open access

[33] http://wikimedia.7.n6.nabble.com/Competition-on-Quality-Flaw-Prediction-in-Wikipedia-PAN-CLEF-12-td4640270.html

[34] Suzuki, Y., & Yoshikawa, M. (2012), QualityRank: assessing quality of wikipedia articles by mutually evaluating editors and texts. *23rd ACM Conference on Hypertext and Social Media* (HT 2012). **DOI** (http://dx.doi.org/10.1145/2309996.2310047) Open access

[35] http://www.wikisym.org/ws2012/bin/view/Main/Program

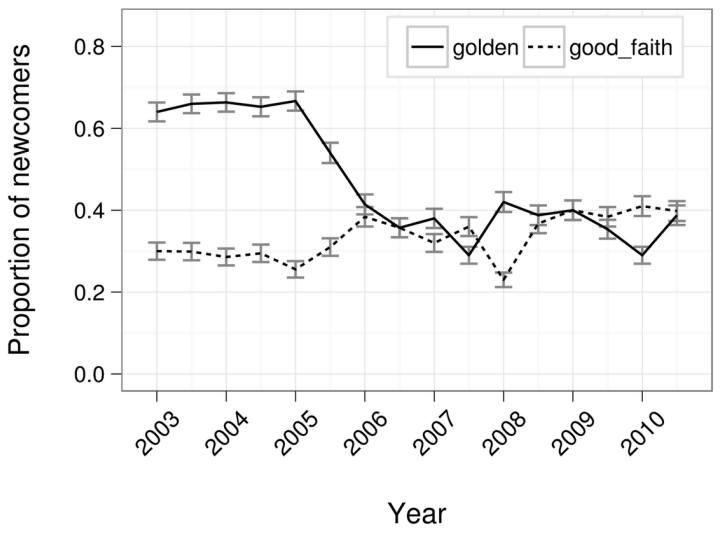
Issue 2(9): September 2012

"Rise and decline" of Wikipedia participation, new literature overviews, a look back at WikiSym 2012

With contributions by: Piotrus, Phoebe, DarTar, Benjamin Mako Hill, Ragesoss and Tbayer

"The rise and decline" of the English Wikipedia

A paper to appear in a special issue of *American Behavioral Scientist* (summarized in the research index) sheds new light on the English Wikipedia's declining editor growth and retention trends. The paper describes how "several changes that the Wikipedia community made to manage quality and consistency in the face of a massive growth in participation have lead to a more restrictive environment for newcomers".[1] The number of active Wikipedia editors has been declining since 2007 and research examining data up to September 2009[2] has shown that the root of the problem has been the declining retention of new editors. The authors show this decline is mainly due to a decline among desirable, good-faith newcomers, and point to three factors contributing to the increasingly "restrictive environment" they face.

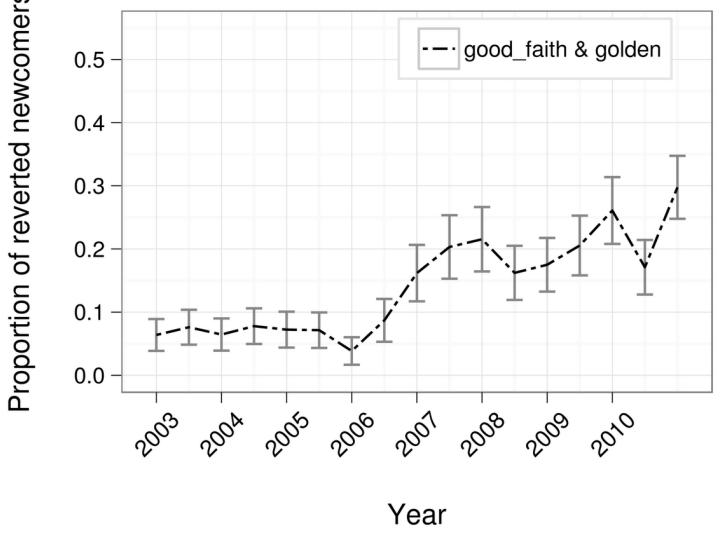


Quality of newcomers over time. The proportion of newcomers falling into the two desirable quality classes is plotted over time.

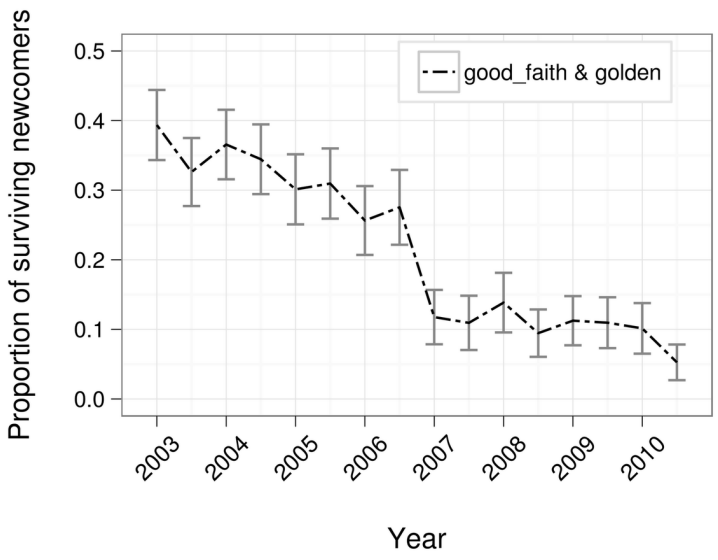
First, Wikipedia is increasingly likely to reject desirable newcomers' contributions, be it in the form of reverts or deletions. Second, it is increasingly likely to greet them with impersonal messages; the authors cite a study that shows that by mid 2008 over half of new users received their first message in a depersonalized format, usually as a warning from a bot, or an editor using a semi-automated tool[3]. They show a correlation between the growing use of various depersonalized tools for dealing with newcomers, and the dropping retention of newcomers. The authors speculate that unwanted but good faithed contributions were likely handled differently in the early years of the project – unwanted changes were fixed and non-notable articles were merged. Startlingly, the authors find that a significant number of first time editors will make an inquiry about their reverted edit on the talk page of the article they were reverted on only to be ignored by the Wikipedians who reverted them. Specifically editors who

use vandal-fighting tools like Huggle or Twinkle are increasingly less likely to follow the Wikipedia:Bold, revert, discuss cycle and respond to discussions about their reverts.

As a third factor, the authors note that the majority of Wikipedia rules were created before 2007 and have not changed much since, and thus new editors face the environment where they have little influence on the rules that govern their behavior, and more importantly, how others should be behave toward them. The authors note that this violates Ostrom's 3rd principle for stable local common pool resource management, by effectively excluding a group that is very vulnerable to certain rules from being able to effectively influence them.



Rejection of desirable newcomers. The proportion of reverted desirable newcomers is plotted over time.



Survival of desirable newcomers. The proportion of surviving desirable newcomers is plotted over time.

The authors recognize that automated tools and extensive rules are needed to deal with vandalism and manage a complex project, but they caution that the currently evolved customs and procedures are not sustainable for the long term. They suggest Wikipedia editors could copy the strategy of distributed, automated tools that have proven so effective at dealing with vandalism (e.g. Huggle & User:ClueBot NG) to build tools that aid in identifying and supporting desirable newcomers (a task in which Wikipedia increasingly fails[4]). Further, they recommend that the newcomers are given a voice, if indirectly via mentors, when it comes to how rules are created and applied.

Overall, the authors present a series of very compelling arguments, and the only complaint this reviewer has is that (even though three of the four were among the Wikimedia Foundation's visiting researchers for the Summer of Research 2011) they do not discuss the fact that the Foundation and the wider community has recognized similar issues, and has engaged in debates, studies, pilot programs and such aimed to remedy the issue (see for example the WMF Editor Trends Study).

Literature reviews of Wikipedia's inputs, processes, and outputs

Nicolas Jullien's "What we know about Wikipedia. A review of the literature analyzing the project(s)"[5] is an attempt at a "comprehensive" literature review of academic research on Wikipedia. Jullien works to distinguish his literature review from previous attempts like those of Okoli and collaborators (cf. earlier coverage: "A systematic review of the Wikipedia literature") and of Park which tend to split the literature into three main themes: (1) motivations of editors to contribute and relationship between motivation and contribution quality, (2) editorial processes and organization and its relationship to quality and (3) the quality and reliability of production.

Jullien builds on this basic framework by Carillo and Okoli, but distinguishes his from their work in several ways. First, Jullien holds that previous work has focused too little on the outputs, which his analysis emphasizes more. Second and crucially, Jullien's review is not limited to material published in journals and, as a result, is more representative of fields like computer science, HCI, and CSCW, which publish many of their most influential articles in conference proceedings. Jullien does not consider articles on how Wikipedia is used, questions of tools and their improvement, and studies that only use Wikipedia as a database (e.g., to test an algorithm). Other than this, the study is not limited to any particular field. It covers articles published in English, French and Spanish before December 2011, mostly based on searches in WebofScience and Scopus (sharing the search query used in the latter). The review is structured around inputs, processes, and outputs.

In terms of inputs, Jullien considers broad cultural factors in the broader environment and questions of why people choose to participate or join Wikipedia. In terms of process, he considers questions about the activities and roles of contributors, the social (e.g., network) structure of both the projects and the individuals who participants, the role of teams and organization of people within them, the processes around editing, creation, deletion, and promotion of articles with a particular focus on conflict, and questions of management and leadership. In terms of outputs, the paper divides publications into studies of process, Wikipedia user experience, the external evaluation of Wikipedia articles, and questions of Wikipedia coverage.

A second recent preprint by Taha Yasseri and János Kertész [6] likewise gives an overview of vast areas of recent research about Wikipedia. Subtitled "Sociophysical studies of Wikipedia" and citing 114 references, it compares some of the authors' own results on e.g. editing patterns (covered in several past issues of this research report, e.g.: "Dynamics of edit wars") with existing literature. The review focuses on quantitative data-driven analyses of Wikipedia production, reproduces and reports a series of previous analyses, and extends some of the earlier findings.

After a detailed description of how Wikipedia works, the authors walk through a series of types of quantitative analyses of patterns of editing to Wikipedia. They use "blocking" of edits to characterize good and "bad" editors and describe different editing patterns between these groups. The authors show that editors, in general, tend to edit in a "bursty" pattern with long periods of breaks and that editing tends to follow daily and weekly patterns that vary by culture. They also walk through several approaches for classifying edits by type, and discuss the characterization of linguistic features with an emphasis on readability.

Much of their article is focused on the issue of conflicts and edit warring. The authors pay particular attention both to the identification of conflicts and of controversial articles and topics, and to characterizing the nature of edit warring itself. The paper ends with the description of an agent-based model of edit warring and conflict.

WikiSym 2012: overview report

The International Symposium on Wikis and Open Collaboration -– "WikiSym 2012" – was held August 27–29 in Linz, Austria. The three-day conference featured research papers, posters and demonstrations, and open space discussion sessions. About 80 researchers and wiki experts from around the world attended.

WikiSym is an academic conference, now in its eighth year, that seeks to highlight research on wikis and open collaboration systems. This year's WikiSym had a strong focus on Wikipedia research, with studies that ranged from analyzing breaking news articles on Wikipedia to looking at the behavior of Wikipedia editors and how long they stay active. In all, 17 papers focused on Wikipedia or MediaWiki, and the two keynotes also focused on Wikipedia research.

The first keynote session was given by Jimmy Wales, who discussed challenges for Wikipedia and potential research questions that matter to the Wikimedia community [7][8]; Wales focused particularly on questions around diversity of the editing body, how to grow small language communities, and how to retain editors. The closing keynote was given by Brent Hecht [9], a researcher from Northwestern University, who spoke on techniques for making multilingual comparisons of content across Wikipedia versions, which in turn allows researchers to identify the potential cultural biases of various Wikipedia editions. Hecht found, for instance, that (looking at interwiki links across 25 languages) the majority of Wikipedia article topics only appear in 1 language; that the overlap between major language editions is relatively small; and that the depth of geographical representation varies widely by language, which a bias towards representing the country or place where that edition's language is prominent. Hecht also compared articles on the same topic across Wikipedias to see the degree of similarity between them. Hecht described his work as "hyperlingual", developing techniques to gain a broader perspective on Wikipedia by looking across language editions. His content comparison tool can be seen at the Omnipedia site [10], and the WikAPIdia API software he developed can be downloaded here [11]. (See also earlier coverage about Omnipedia: "Navigating conceptual maps of Wikipedia language editions")

In addition to the presented papers, some of which are profiled below, WikiSym has a strong tradition of hosting open space sessions in parallel with the main presentations, so that attendees can discuss topics of interest. This year's open space topics included helping new wiki users; non-text content in wikis (including videos, images, annotations, slideshows and slidecasting); the future of WikiSym; Wikipedia bots; surveying Wikipedia editors; and realtime wiki synchronization and multilingual synchronization feedback. The conference closed with a panel session entitled "What Aren't We Measuring?", where panelists discussed and debated various methods for quantifying wiki-work (by studying editors, edits, and other metrics).

This year's WikiSym was hosted at the Ars Electronica Center in Linz, a "museum of the future" that hosts the Ars Electronica festival every year. The colorful, dramatic Ars Electronica building is in the heart of Linz, so outside of sessions conference attendees enjoyed exploring and socializing in the city center. The conference dinner was held at the Pöstlingberg Schlössl, which is accessed by one of the steepest mountain trams in the world.

WikiSym 2012 papers and poster and demonstration abstracts may be downloaded from the conference website [12]. Next year's WikiSym is planned for Hong Kong, just before Wikimania 2013. Updates on the schedule and important dates can be found on the WikiSym blog [13].

On the "Ethnography Matters" blog, participant Heather Ford looked back at the conference,[14] stating that "WikiSym is dominated by big data quantitative analyses of English Wikipedia", asking "where does ethnography belong?" and counting 82% of the Wikipedia-related papers as examining the English Wikipedia and only 18% about other language Wikipedias. A panel at WikiSym 2011 had called to broaden research to other languages (see last year's coverage: "Wiki research beyond the English Wikipedia at WikiSym").

WikiSym 2012 papers

The conference papers and posters [12] included, (apart from several ones that have been covered in earlier issues of this report):

- **{{Citation needed}}: The dynamics of referencing in Wikipedia**[15]: This paper contributes to the debates on Wikipedia's reliability. The authors find that density of references is correlated with the article length (the longer the article, the more references it will have per given amount of text). They also find that references attract more references (suggesting a form of a snowball mechanism at work) and that the majority of references is added in short periods of time by editors who are more experienced, and who are also adding substantial content. The authors thus conclude that referencing is primarily done by a small number of experienced editors, who prefer to work on longer articles, and who drastically raise the article's quality, by both adding more content, and by adding more references.
- **Etiquette in Wikipedia: Weening** [*sic*] **New Editors into Productive Ones**[16]: The authors of this paper experimented with alternative warning messages, introducing a set of shorter and more personalized warnings into those delivered by Huggle in the period of November 8 0 December 9 2011. Unfortunately, the authors are rather unclear on how exactly the Huggle tool was influenced, and whether the community was consulted on that. While in fact the community and Huggle developers have been aware of, discussed and approved of this experiment – here or here – the paper's omission to clarify that this was the case can lead to some confusion with regard to research ethics, since a casual reader may assume the researchers have hijacked Huggle without consulting the community. The wording changes were in good faith (making the messages more personalized, friendly and short), and the authors conclude that the new messages they tested proved more conductive to positively influenced new editors who received Level 1 Warnings.
- **WikiTrust algorithm applied to MediaWiki programmers**: A paper titled "Towards Content-driven Reputation for Collaborative Code Repositories"[17] reports on an experimental application of the well-known WikiTrust algorithm to the collaboration of programmers on a code repository, namely MediaWiki's own SVN codebase (from 2011, before it was switched to Git). In that model, contributors lose reputation when their contributions are reverted or deleted. According to the abstract, "Analysis is particularly attentive to reputation loss events and attempts to establish ground truth using commit comments and bug tracking. A proof-of-concept evaluation suggests the technique is promising (about two-thirds of reputation loss is justified) with false positives identifying areas for future refinement." An example of such false positives is "The “not now” trap: Frequently a change is reverted with a 'not now' justification, e.g., needing to hold for more testing. When that testing is done the changes are likely to be re-committed in much the same form, punishing the benign reverting editor."
- **"Deletion Discussions in Wikipedia: Decision Factors and Outcomes"**[18] found among other things that "69.5% of discussions and 91% of comments are well-represented by just four factors: Notability, Sources, Maintenance and Bias. The best way to avoid deletion is for readers to understand these criteria." One of the authors also co-presented a demo showing mock-ups of possible "alternative interfaces for deletion discussions in Wikipedia"[19], which would highlight the prevalence of each type of argument (e.g. notability, sourcing...) in a deletion discussion more clearly.
- **"Classifying Wikipedia articles using network motif counts and ratios"**[20]: Similar to an earlier paper by the same authors (earlier coverage: "Collaboration pattern analysis: Editor experience more important than 'many eyes'"), this paper examined the collaboration network of Wikipedia articles and editors using Network motifs – 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. This was then related to the quality ratings of articles: "Pages with good quality scores [e.g. featured articles] have characteristic motif profiles, but pages with good user ratings [from the Article Feedback tool] don’t. This suggests that a good quality score is evidence that a collaborative curation process has been pursued. However, not all pages with high quality scores get good user ratings and some pages with low quality scores are trusted by users. Perhaps the Wikipedia quality scale is a low error scale rather than a

quality scale?"

- **"'Writing up rather than writing down': becoming Wikipedia Literate"**[21] applied "the work of literacy practitioner and theorist Richard Darville" to communication among Wikipedians, e.g. new users and experienced users who deleted some of their contributions. "Using a series of examples drawn from interviews with new editors and qualitative studies of controversies in Wikipedia, we identify and outline several different literacy asymmetries."
- **"How long do wikipedia editors keep active?"**[22] found that on the English Wikipedia, "although the survival function of occasional editors roughly follows a lognormal distribution, the survival function of customary editors can be better described by a Weibull distribution (with the median lifetime of about 53 days). Furthermore, for customary editors, there are two critical phases (0–2 weeks and 8–20 weeks) when the hazard rate of becoming inactive increases".

"First Monday" on rhetoric, readability and teaching

First Monday, the veteran open access journal about Internet topics, featured three Wikipedia-themed papers in its September issue:

- **AfD rhetoric examined**: "The pentad of cruft: A taxonomy of rhetoric used by Wikipedia editors based on the dramatism of Kenneth Burke"[23] is an essay "describing a method for classifying arguments made by Wikipedia editors based on the theory of 'dramatism,' developed by the literary theorist Kenneth Burke, and demonstrating how this method can be applied to a small sample of arguments drawn from Wikipedia's 'Article for Deletion' (AfD) process."
- **"Readability of Wikipedia"**[24] applied the standard Flesch Reading Ease test to the English and Simple English Wikipedias (at http://www.readabilityofwikipedia.com/, the authors also offer the possibility to view scores directly). The effort, described as "extensive research" in an university press release [25] found that "overall readability is poor, with 75 percent of all articles scoring below the desired readability score. The 'Simple English' Wikipedia scores better, but its readability is still insufficient for its target audience." See also the detailed earlier *Signpost* coverage: "Readability of Simple English and English Wikipedias called into question", and the summary of an earlier paper which applied a more diverse set of readability measures to both Wikipedias: "Simple English Wikipedia is only partially simpler/controversy reduces complexity"
- **"Wikis and Wikipedia as a teaching tool: Five years later"**[26] by longtime Wikipedian (and contributor to this research newsletter) Piotr Konieczny first gives an overview over the now widespread use of Wikipedia in the classroom and its advantages, and in a second part offers detailed practical advice drawing from the author's own "five years of experience in teaching with wikis and Wikipedia and holding workshops on the subject".

Briefly

- **Recent changes visualization designed to assist admins**: A paper titled "Feeling the Pulse of a Wiki: Visualization of Recent Changes in Wikipedia"[27] will be presented at the upcoming conference "VINCI 2012 : The International Symposium on Visual Information Communication and Interaction". It describes a prototype software (apparently not publicly available yet) that is designed "to aid a wiki administrator to perceive current activity in a wiki", starting out from the idea to map editors and articles in two dimensions: time and activity level. Hosted on the Toolserver, the software directly accesses a wiki's Recent Changes table, containing edits from the last 30 days. Using their tool, the authors visually discerned "six common editing patterns" on the English Wikipedia. E.g. "*New article, many editors, many edits*: this is the *new popular article* pattern which almost invariably reflects a current event". The authors also compare their tool to the previous "few and limited efforts" to visualize recent changes: WikipediaVision [28], Wikipulse [29] and Wikistream [30].
- **Unearthing the "actual" revision history of a Wikipedia article**: A paper[31] by two researchers from Waseda University observes that "Unlike what is very common in software development, Wikipedia does not maintain an

explicit revision control system that manages the detailed change through revisions. The chronologically-organized edit history fails to reveal the meaningful scenarios in the actual evolution process of Wiki articles, including reverts, merges, vandalism and edit wars". To extract this "actual" revision graph, where two neighboring nodes correspond to a revision and an earlier one which it was derived from, a similarity measure is needed. The article cites a 2007 paper [32] and other research which had already proposed to understand a page's revision history as a directed tree and used similarity measures such as tf-idf. The present paper uses a similarity measure based on the frequency of n-grams (sequences of *n* words) and goes further in regarding the revision history as a directed acyclic graph. This allows for version merges, although the actual algorithm presented still focuses on the case of trees.

- **Who deletes Wikipedia – or reverts it**: Wibidata, a big data analytics startup based in San Francisco, posted a follow-up[33] to their "Who deletes Wikipedia" analysis (previous coverage), taking into account the effect of reverts, which several Wikipedians had pointed out in response to their earlier blog post.
- **Geospatial characteristics of Wikipedia articles**: The authors of this paper attempt to identify what makes Wikipedia articles with geographical coordinates different from others (besides their obvious relation to geographical locations).[34] They rather unsurprisingly find that more developed articles are more likely to have geo-coordinates, and consequently they find that there seems to be a correlation between article quality and having geo-coordinates links. They also find that articles with geo-coordinates are more likely to be linked to, a likely function of them being of above-average quality.
- **Wikipedia's affordances**: This paper, framing itself as part of the ecological psychology field, contribute to the discourse about affordances (property of an object that allows one to take a certain action).[35] The authors submit that this concept can be developed to further our understanding of how individuals perceive their socio-technical environment. The authors refine the term "technology affordances", which they define as "functional and relational properties of the user-technology system". Then use Wikipedia as their case study attempting to demonstrate its value, listing six affordances of Wikipedia (or in other words, they note that editors of Wikipedia can take the following six actions): contribution, control, management, collaboration, self-presentation, broadcasting.
- **Hematologists unsure whether "to engage with Wikipedia more constructively"**: A letter[36] to the medical journal BMJ asks "Should clinicians edit Wikipedia to engage a wider world web?" The authors, a student and a senior lecturer in the field of haematology, "simulated 30 opportunistic internet searches for information on haemophilia in the top three search engines using term permutations: haemophilia or hemophilia (with or without A or B); carrier; information; child; treatment. Wikipedia was the most commonly found top 10 site in all search engines." In an apparent attempt to gauge the authoritativeness of Wikipedia content, "Analysis of editorial authorship of the Haemophilia Wiki [sic] for four weeks found 39 edits by 25 editors, only nine of whom had a profile, and none of whom were experts in haemophilia." Possibly unaware of Wikipedia's "no original research" policy, the authors ask "Given the evolving debate about open access to data, should publishers and authors be mandated to place reviews and key studies [...] in a public domain like Wikipedia?" (naming the example of a recent prominent paper in the field, which the Wikipedia article cites only in form of a New York Times news article about it). The letter concludes "as a professional group, we are not sure whether we wish to engage with Wikipedia more constructively". One-day access to the letter, which is around half a page long, can be purchased at £20/$30/€32 plus VAT, which may not be a very competitive price given the availability of more thorough evaluations of Wikipedia's quality elsewhere in the academic literature.
- **Tracking and verifying sources on Wikipedia**: Ethnographer Heather Ford published the final report from her study on how editors track and verify sources on Wikipedia.[37] The report presents an in-depth qualitative analysis of editor discussions around verifiability of information in the early editing phase of the 2011–2012 Egyptian revolution article and reviews how Wikipedia policies around primary vs secondary sources, notability and neutrality were used to make decisions about what sources to cite.

- **A recommender system for infoboxes**: A team of computer science researchers at the University of Texas at Arlington developed a classification method to predict infobox template types from articles lacking them, using three types of features: words in articles, categories, and named entities (or words with corresponding Wikipedia entries). The study suggests that articles with infoboxes and articles without infoboxes exhibit a substantially different distributions of the above features. The classifier was tested on data from a 2008 dump of the English Wikipedia.[38]
- **Styles of information search on Wikipedia**: A poster presented at the *2nd European Workshop on Human-Computer Interaction and Information Retrieval* presents the results of an eye-tracking study looking at patterns of information search in Wikipedia articles. The study looks at task-specific differences in the context of factual information lookup, learning and casual reading activity.[39]
- **Post-edit feedback experiment**: The Wikimedia Foundation's "Editor Engagement Experiments" team reported[40] on an experiment with a simple user interface change – adding messages that confirm that an edit has been saved – and its effect on the contributions of new editors.
- **Pilot study about Wikipedia's quality compared to other encyclopedias**: The results of a pilot study commissioned by the Wikimedia Foundation [41], titled "Assessing the Accuracy and Quality of Wikipedia Entries Compared to Popular Online Alternative Encyclopaedias: A Preliminary Comparative Study Across Disciplines in English, Spanish and Arabic"[42] have been announced [43].
- **Wikipedia, the first step toward communism**: Sylvain Firer-Blaess and Christian Fuchs, in their "info-communist manifesto", argue that Wikipedia is an example of the communist mode of production and participatory democracy—"the brightest info-communist star on the Internet's class struggle firmament". They suggest that Wikipedia's future will be a choice between co-option into the broader capitalist economy (through the exploitation of the commercial possibilities of Wikipedia's free licensing) or, alongside similar "info-communist" projects, displacing more and more capitalist production of informational goods.[44]
- **Quality flaw detection competition**: Maintenance templates on the English Wikipedia (e.g. "citation needed") have attracted the attention of several researchers recently, as easy to parse indicators of quality problems (example). An "Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia"[45] summarizes its outcome as follows: "three quality flaw classifiers have been developed, which employ a total of 105 features to quantify the ten most important quality flaws in the English Wikipedia. Two classifiers achieve promising performance for particular flaws. An important 'by-product' of the competition is the first corpus of flawed Wikipedia articles, the PAN Wikipedia quality flaw corpus 2012 (PAN-WQF-12)", which consists of "1 592 226 English Wikipedia articles, of which 208 228 have been tagged to contain one of ten important quality flaws". One of the two "winners", the "FlawFinder" algorithm, has been described in a paper covered last month. The competition took place on occasion of the CLEF 2012 conference, as did the first Wikipedia Vandalism Detection competition [46] two years ago (*Signpost* coverage).

References

[1] Halfaker, A., Geiger, R.S., Morgan, J. and Riedl, J. (2012), The Rise and Decline of an Open Collaboration Community, *American Behavioral Scientist*, forthcoming. **HTML** (http://www-users.cs.umn.edu/~halfak/summaries/The Rise and Decline/) summary Open access

[2] http://strategy.wikimedia.org/wiki/Editor_Trends_Study

[3] Geiger, R. S., Halfaker, A., Pinchuk, M., & Walling, S. (2012). Defense Mechanism or Socialization Tactic? Improving Wikipedia's Notifications to Rejected Contributors. ICWSM.

[4] Musicant, D. R., Ren, Y., Johnson, J. A., & Riedl, J. (2011). Mentoring in Wikipedia: a clash of cultures. WikiSym 2011 (pp. 173–182). (http://www.wikisym.org/ws2011/_media/proceedings:p173-musicant.pdf)

[5] Jullien, N. (2012). What We Know About Wikipedia: A Review of the Literature Analyzing the Project(s). SSRN Electronic Journal. **PDF** (http://papers.ssrn.com/abstract=2053597) Open access

[6] Yasseri, T., & Kertész, J. (2012). Value production in a collaborative environment. Physics and Society; Computers and Society; Data Analysis, Statistics and Probability. **PDF** (http://arxiv.org/abs/1208.5130) Open access

[7] https://twitter.com/dirkriehle/status/240426491547099136/photo/1

[8] https://twitter.com/dirkriehle/status/240427461240844288/photo/1

[9] http://brenthecht.com/
[10] http://omnipedia.northwestern.edu/
[11] http://collablab.northwestern.edu/wikapidia_api/Wikapidia/Home.html
[12] http://wikisym.org/ws2012/bin/view/Main/Program
[13] http://www.wikisym.org/
[14] Ford, H. (2012) Where does ethnography belong? Thoughts on WikiSym 2012, *Ethnography Matters* **HTML** (http://ethnographymatters.net/2012/09/06/where-does-ethnography-belong/) Open access
[15] Chen, C.-C. and Roth, C. (2012), {{Citation needed}}: The dynamics of referencing in Wikipedia, *WikiSym '12* **PDF** (http://www.wikisym.org/ws2012/bin/download/Main/Program/p19wikisym2012.pdf) Open access
[16] Faulkner, R., Walling, S. and Pinchuk, M. (2012), Etiquette in Wikipedia: Weening New Editors into Productive Ones, *WikiSym '12* **PDF** (http://www.wikisym.org/ws2012/bin/download/Main/Program/p17wikisym2012.pdf) Open access
[17] West, A.G. and Lee, I. (2012) Towards Content-driven Reputation for Collaborative Code Repositories, *WikiSym '12* **PDF** (http://www.wikisym.org/ws2012/bin/download/Main/Program/p5wikisym2012.pdf) Open access
[18] Schneider, J., Passant, A. and Decker, S. (2012) Deletion Discussions in Wikipedia: Decision Factors and Outcomes, *WikiSym '12* **PDF** (http://jodischneider.com/pubs/wikisym2012.pdf) Open access
[19] Jodi Schneider, Krystian Samp: Alternative Interfaces for Deletion Discussions in Wikipedia: Some Proposals Using Decision Factors. Demo, WikiSym'12, August 27–29, 2012, Linz, Austria. ACM 978-1-4503-1605-7/12/08. **PDF** (http://wikisym.org/ws2012/bin/download/Main/Program/Schneider.pdf) Open access
[20] Wu, G., Harrigan, M. and Cunningham, P. (2012) Classifying Wikipedia Articles Using Network Motif Counts and Ratios, *WikiSym '12* **PDF** (http://www.wikisym.org/ws2012/bin/download/Main/Program/p2awikisym2012.PDF) Open access
[21] http://wikisym.org/ws2012/bin/download/Main/Program/p21wikisym2012.pdf
[22] http://wikisym.org/ws2012/bin/download/Main/Program/p15wikisym2012.pdf
[23] Famiglietti, Andrew. The pentad of cruft: A taxonomy of rhetoric used by Wikipedia editors based on the dramatism of Kenneth Burke. First Monday [Online], (19 August 2012) **HTML** (http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/4082/3294) Open access
[24] Lucassen, Teun, Dijkstra, Roald, AND Schraagen, Jan Maarten. "Readability of Wikipedia" First Monday[Online], (20 August 2012) **HTML** (http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3916) Open access
[25] http://www.utwente.nl/en/archive/2012/09/research_at_the_university_of_twente_wikipedia_article_readability_too_low.doc/
[26] Konieczny, Piotr. "Wikis and Wikipedia as a teaching tool: Five years later" First Monday [Online], (25 August 2012) **HTML** (http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3583) Open access
[27] Robert P. Biuk-Aghai and Roy Chi Kit Chan (2012) Feeling the Pulse of a Wiki: Visualization of Recent Changes in Wikipedia, *VINCI 2012*, forthcoming **PDF** (http://www.cad.zju.edu.cn/home/chenwei/VINCI2012/content/4.2.pdf) Open access
[28] http://www.lkozma.net/wpv/
[29] http://wikipulse.herokuapp.com/
[30] http://wikistream.inkdroid.org/
[31] Wu, J., & Iwaihara, M. (2012). Wikipedia Revision Graph Extraction Based on N-Gram Cover. In Z. Bao, Y. Gao, Y. Gu, L. Guo, Y. Li, J. Lu, Z. Ren, et al. (Eds.), *Lecture Notes in Computer Science*, 2012, Volume 7419 (Vol. 7419, pp. 29–38). Berlin, Heidelberg: Springer Berlin Heidelberg. **DOI** (http://dx.doi.org/10.1007/978-3-642-33050-6) Closed access
[32] http://www.wikisym.org/ws2007/_publish/Sabel_WikiSym2007_StructuringRevision.pdf
[33] Hougland, J. (2012) Reverting in Wikipedia, *Wibidata blog* **HTML** (http://www.wibidata.com/2012/09/17/reverting-in-wikipedia/) Open access
[34] Hahmann, S. and Burghardt, D. (2012), *Investigation on factors that influence the (geo)spatial characteristics of Wikipedia articles* **PDF** (http://giscience.org/proceedings/abstracts/giscience2012_paper_84.pdf) Open access
[35] Mesgari, M. and Faraj, S. (2012) Technology Affordances: The Case of Wikipedia, *AMCIS 2012* **PDF** (http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1303&context=amcis2012) Open access
[36] Kint, M., & Hart, D. P. (2012). Should clinicians edit Wikipedia to engage a wider world web? BMJ (Clinical research ed.), 345, e4275. **PDF** (http://www.ncbi.nlm.nih.gov/pubmed/22761090) Closed access
[37] Ford, H. (2012) Wikipedia Sources: Managing Sources in Rapidly Evolving Global News Articles on the English Wikipedia, *SSRN*, August 2012. **PDF** (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2127204) Open access
[38] Sultana, A., Hasan, Q.M.,, Biswas, A.K., Das, S., Rahman, H., Ding, C. and Li, C. (2012), Infobox Suggestion for Wikipedia Entities, *21st ACM International Conference on Information and Knowledge Management* (CIKM '12) **PDF** (http://ranger.uta.edu/~cli/pubs/wikiclassification-cikm12poster-final-shbdrdl-aug12.pdf) Open access
[39] Knäusl, H., Elsweiler, D. and Ludwig, B. (2012) Towards Detecting Wikipedia Task Contexts, *2nd European Workshop on Human-Computer Interaction and Information Retrieval*, August 2012 **PDF** (http://red.cs.nott.ac.uk/~mlw/EuroHCIR2012/poster4.pdf) Open access
[40] Walling, S. and Taraborelli, D. (2012), Is this thing on? Giving new Wikipedians feedback post-edit, *Wikimedia Blog* **HTML** (https://blog.wikimedia.org/2012/09/24/giving-new-wikipedians-feedback-post-edit/) Open access
[41] https://blog.wikimedia.org/2012/08/02/seven-years-after-nature-pilot-study-compares-wikipedia-favorably-to-other-encyclopedias-in-three-languages/

[42] Casebourne, I., Davies, C., Fernandes, M., Norman, N. (2012): *Assessing the Accuracy and Quality of Wikipedia Entries Compared to Popular Online Alternative Encyclopaedias: A Preliminary Comparative Study Across Disciplines in English, Spanish and Arabic*. **PDF** (https://commons.wikimedia.org/wiki/File:EPIC_Oxford_report.pdf) Open access

[43] http://epiclearninggroup.com/uk/news/pilot-comparative-study-of-online-encyclopaedias-yields-insights-into-wikipedias-accuracy-and-quality/

[44] Sylvain Firer-Blaess and Christian Fuchs (2012), Wikipedia: An Info-Communist Manifesto, *Television & New Media*, 12 September 2012 abstract (http://tvn.sagepub.com/content/early/2012/09/10/1527476412450193) Closed access

[45] Maik Anderka and Benno Stein: Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In: Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.): *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, 17–20 September, Rome, Italy. ISBN 978-88-904810-3-1. ISSN 2038-4963. 2012. **PDF** (http://www.uni-weimar.de/medien/webis/publications/papers/stein_2012u.pdf) Open access

[46] http://web.archive.org/web/20101011035833/http://www.uni-weimar.de/medien/webis/publications/downloads/papers/stein_2010t.pdf

Issue 2(10): October 2012

WP governance informal; community as social network; efficiency of recruitment and content production; Rorschach news

With contributions by: Piotrus, Adler.fa, Bdamokos, Ragesoss, Tbayer, and Phoebe

Wikipedia governance found to be mostly informal

A paper in the *Journal of the American Society for Information Science and Technology*, coming from the social control perspective and employing the repertory grid technique, has contributed interesting observations about the governance of Wikipedia.[1] The paper begins with a helpful if cursory overview of governance theories, moving towards the governance of open source communities and Wikipedia. That cursory treatment is not foolproof, though: for example, the authors mention "bazaar style governance", but attribute it incorrectly—rather than the 2006 work they cite, the coining of this term dates to Eric S. Raymond's 1999 *The Cathedral and the Bazaar*. The authors have interviewed a number of Wikipedians and identified a number of formal and informal governance mechanisms. Only one formal mechanism was found important—the policies—while seven informal mechanisms were deemed important: collaboration among users, discussions on article talk pages, facilitation by experienced users, individuals acting as guardians of the articles, inviting individuals to participate, large numbers of editors, and participation by highly reputable users. Notably, the interviewed editors did not view elements such as administrator involvement, mediation or voting as important.

The paper concludes that "in the everyday practice of content creation, the informal mechanisms appear to be significantly more important than the formal mechanisms", and note that this likely means that the formal mechanisms are used much more sparingly than informal ones, most likely only in the small percentage of cases where the informal mechanisms fail to provide an agreeable solution for all the parties. It was stressed that not all editors are equal, and certain editors (and groups) have much more power than others, a fact that is quickly recognized by all editors. The authors note the importance of transparent interactions in spaces like talk pages, and note that "the reported use of interaction channels outside the Wikipedia platform (e.g., e-mail) is a cause for concern, as these channels limit involvement and reduce transparency." Citing Ostrom's governance principles, they note that "ensuring participation and transparency is crucial for maintaining the stability of self-governing communities."

Social network analysis of Wikipedia community

This paper looks at the relationships between Wikipedians from the social network analysis perspective (nodes are defined as authors, and links as indicators of collaboration on the same article), treating Wikipedia as an online social network (similar to Facebook).[2] The authors note that while Wikipedia is not *primarily* a social network site, it has enough social networking qualities to justify being seen as such. They find that Wikipedia can be seen as a very good source of information about online relationships between actors, due to the transparent and public nature of its data. The authors present a brief overview of previous work with a similar approach. Rather unsurprisingly, the authors find that in the very early days of Wikipedia, editors were much more likely to know one another and collaborate on articles than in the later years. They find that the number of editors is highly correlated to the editors' familiarity with one another, and is more relevant than the number of articles, as they find that from 2007, when the number of editors roughly stabilized, so did their levels of connectedness through collaboration.

The paper shows that with very few exceptions (low activity, specialized editors) all Wikipedia editors are connected to one another, and there are no isolated groups (or topic areas). The authors also find that the Wikipedia collaborations can be analyzed using the small-world network approach (suggesting that the distance between editors, defined as the average path length, with links being articles contributed to, is very small). The article focuses primarily on the mathematical side of social network analysis, and unfortunately offers little commentary or analysis of the findings. The validity of the results can also be questioned, as the authors treat bots and semi-automated accounts as "regular authors"; considering that the majority of Wikipedia articles have been edited by bots or editors using scripts, the finding that editor A can be connected to editor B through the fact that they both edited different pages which in turn were edited by the same bot or script-equipped editor is hardly surprising.

Wikipedia's article on the Rorschach inkblot test found to have a limited effect on the test's results

One of the inkblot images from the original Rorschach test, illustrating the Wikipedia article about it (here shown without the list of popular responses that according to Schultz and Brabender appears to influence test results)

Earlier this month, the *Journal of Personality Assessment* published a paper titled "More Challenges Since Wikipedia: The Effects of Exposure to Internet Information About the Rorschach on Selected Comprehensive System Variables".[3] Summarizing past events (well-known to Wikipedians) from the point of view of psychologists adhering to the Rorschach test as a diagnostic tool, they write: "The availability of Rorschach information online has become of even greater concern in the last few years, since James Heilman, an emergency-room physician from Canada, posted images of all 10 Rorschach inkblots on the popular online encyclopedia, Wikipedia (Cohen, 2009; Wikipedia, 2004[sic]). This Wikipedia article also describes “common responses” to each blot, which frequently correspond to percepts that would be scored Popular under the current coding rules of Exner’s (2003) Comprehensive System (CS)." They remark that "Although many psychologists decried the publishing of the Rorschach inkblots on Wikipedia, before this study, no published studies had examined whether viewing the inkblots and other Rorschach information posted on Wikipedia would impact examinees’ scores." (As reported last year in this newsletter - see "Psychologists gauge impact of Wikipedia's Rorschach test coverage" - one of the authors had coauthored a study that had investigated the rise in prominence of information about the test on the Internet due to Wikipedia, but not tested its impact on the test itself.)

Before reporting their own results, the authors cite an unpublished dissertation,[4] which had compared test subjects' Rorschach results before and after reading the article. Its tentative results suggested a "significant increase in shading responses [which] then likely affected the corresponding increase in [one variable], but otherwise indicated "that the

majority of CS variables do not appear to be affected by exposure to information in the Wikipedia article."

The authors' own study involved 50 participants, half of whom had to read an excerpt of the Rorschach test article (while the control group read one of the Philadelphia Phillies article) before trying to "fake good" on the test, impersonating a character which would have a huge incentive to achieve certain results in the test ("Jack is a 35-year-old father of two wonderful children ...The judge ordered that Jack have a psychological evaluation done to determine whether or not he should be given custody of his kids.")

Among the test features defined in the "CS" system, only "Populars" was found to differ significantly "between the control and experimental groups [...] likely due to the fact that the Rorschach [Wikipedia article excerpt] provided pictures of each of the inkblots, along with "common responses," which, in many cases, corresponded to those responses that are actually coded as Popular according to the CS. However, the Wikipedia information on its own did not appear to directly impact other variables associated with perceptual accuracy."

Commenting on the paper, Heilman told this research report:

> *That reading about the Rorschach before testing affects scores in a group of "normal" individuals is not really surprising. This analysis, however, does not show that the availability of information regarding psychological tests affects clinical important outcomes.*

Efficiency of Wikipedia in editor recruitment and content production

A paper titled "Is Wikipedia Inefficient? Modelling Effort and Participation in Wikipedia"[5] will be presented at next year's HICSS '13 conference. The main research concern of the authors is whether the saturation observed in the growth of Wikipedia is due to the maturity of the project or is rather caused by editorial obstacles and inefficient collaboration processes. To address this question, they try to investigate the efficiency of collaboration in 39 language editions of Wikipedia. Two different processes are studied. 1) editor recruitment; the ability of Wikipedia projects to attract editors from the pool of potential editors and 2) the article creation process. For each of these two processes corresponding input and output parameters are chosen and by applying a set of Data Envelopment Analysis the relative efficiency of language projects is calculated. For the editor recruitment process the input parameter is the size of the population speaking the language, having access to Internet and being at a tertiary-level of education and the output is the number of Wikipedia editors contributing to the Wikipedia edition of that language. It is shown that the efficiency of some language editions, e.g. Estonian, Hungarian, Norwegian, and Finnish, are much higher than some other language editions, e.g., Malaysian, Arabic, and Chinese. A decreasing return to scale is reported for all of the studied projects; however, the effect is more pronounced for larger ones. In other words, larger projects can be considered as inefficient in attracting new editors. For the production process, the number of Wikipedia editors is considered this time as the input and 3 outputs: number of edits, number of articles, and number of Featured articles. Here, the results generally suggest that for the larger projects the returns to scale are systematically decreasing, showing the difficulties of maintaining the efficiency of the workflow as the project grows. Some projects, such as the Malaysian and Persian Wikipedias, are not as successful in editor recruitment but are still efficient in creating articles given the capacity of their human resources. As for the quality of articles, it is shown that in larger projects like French and German, the focus is more on increasing the quality of the existing articles, whereas in intermediate-size projects, e.g., Russian and Italian, the main effort is still on increasing the number of articles.

The paper notes a positive correlation between efficiency in the number of edits and the efficiency in number of articles and featured articles. Among the limitations of the study, the authors name the time period of the analysed data, being limited to one month, and the possible flaws in the demographic data used to estimate the input of the editor recruitment process. Excluding contributions from unregistered users due to technical reasons could also have induced biases in the results. Since the article starts by raising the question of efficiency of Wikipedia in general, it ends up by comparing different language editions to each other and presenting the results in only relative terms. The English Wikipedia, which could be a benchmark for such comparisons, is entirely excluded from the study. More

importantly, applying the data envelopment analysis, which is originally introduced for evaluating activities of not-for-profit entities participating in public programs, on Wikipedia activity data is not well justified.

Student use of Wikipedia

How students find and evaluate information is a perpetual concern for librarians, who act as educators and guides to finding the best resources for student information needs as well as collection curators. Since the arrival of Wikipedia, librarians have grappled with how the site fits in with and compares to a more traditionally published and reviewed collection, and how best to help students understand and use Wikipedia. This study is an up-to-date addition to the body of literature on this subject.[6] Colón-Aguirre and Fleming-May use a coded qualitative interview approach to understanding undergraduate opinions about Wikipedia, compared to their use of and attitude towards traditional library resources.

The authors conducted interviews with 21 undergraduate students in one college in a large public university in the United States. Based on student responses about their research habits, the authors divided their respondents into three categories: avid library users, occasional library users, and library avoiders. While all categories of students used Wikipedia, there were differences in purpose; avid library users used Wikipedia to gather background information before turning to library-supplied resources like books and journals, while library avoiders relied more on Wikipedia and were lost if they could not find the information they needed on the site or via Google searches. Most of the students interviewed reported getting to Wikipedia via Google or other search engines, and the authors do not report any deep awareness by the students of how the site works or how to evaluate articles; awareness of ability to contribute was not mentioned. Student use of the library versus Wikipedia was also influenced by their perceptions of library resources being difficult to use (both in-person stacks and subscription online resources), particularly compared to the ease of using Wikipedia and online searching; students were also swayed in whether they used the library by their assignment requirements and faculty advice, including professors who advised against using Wikipedia as being "not credible" and required using library resources specifically.

The authors conclude that librarians need to work more with teaching faculty to craft research assignments, and that hands-on instruction in the use of the library does aid student comfort with research. This short article will be most of interest to practicing librarians and undergraduate instructors, who will doubtless see reflections of their own students in the student interviews. Wikipedians who are involved in academic classroom education and outreach will also find this study interesting, if for no other reason than to reinforce the importance of helping students become more knowledgeable about the ways that Wikipedia works with and differs from traditional academic publications.

In brief

- **"Conflict positively influences group performance"**: Investigating the question "Does conflict matter in the success of mass collaboration?", a paper in the Chinese Journal of Library and Information Science[7] investigates conflict on Wikipedia, analyzing it from the social network analysis perspective (nodes are defined as individuals, and links, as indicators of conflict), and differentiating between positive and negative types of conflict. Their goal is to increase understanding of the conflict mechanism in the mass collaboration setting. The authors find that "that participation positively influences task complexity, conflict, and group performance; task complexity positively influences group performance but negatively influences conflict; and conflict positively influences group performance".
- **Generating a lexical network from Wiktionary**: [8] The researcher has created an open source tool – available at http://dbnary.forge.imag.fr/ – that extracts a lexical network (including definitions, translations, synonyms, antonyms, etc.) from Wiktionary data in RDF format, that can be used in existing semantic tools. The author notes that because Wiktionary – unlike traditional dictionaries – treats homonyms (words that share the same spelling and pronunciation but have a different meaning) on single pages with multiple etymology sections, it has not been possible to properly attribute the senses and lexical relations to the proper etymologies (i.e. lexemes).

- **How are article edits and page views related? We still don't know.**: [9] This paper attempts to explore the relationship between the "production" and "consumption" of Wikipedia content: the edits that build articles, and the page views from readers. For broad topic areas on English Wikipedia (such as articles in Category:Dance and its subcategories), the pattern of edits mirrors the overall trend of editing activity—rising exponentially until peaking around 2007, with a linear decline in edit rate since then. Page views for these topic areas, by contrast, show an approximately linear rise page views since late 2007 (which is the earliest period for which we have article traffic statistics [10]). According to the authors, this pattern "conforms to a two-phase evolution framework: one of production followed by consumption", although they do not attempt to establish a causal link between the article content maturation and readership. Unfortunately, the lack of earlier data on article traffic makes it hard to learn much from the relationship between edit rate and article traffic, without taking a more fine-grained approach to identify articles or topic areas whose early phases of rising and peaking edit rates are also covered by page view data.
- **More WikiSym reports**: Two more reports from August's annual WikiSym conference were published this month, by the recipient of a travel grant from the UK Wikimedia chapter,[11] and by a Natural Language Processing (NLP) researcher[12] who dubbed the conference "WikipediaSym" because "the conference submissions were mostly inclined towards the information analysis and social aspects of using wikis, in particular Wikipedia, and there were very few submissions on the actual *applications* of wikis (or wiki-like systems) and the *open collaboration* context". (See also the overview report in the last issue of the research report)
- **German centrality**: A discussion paper examined "Centrality and Content Creation in Networks [in] The Case of German Wikipedia". [13]
- **Systemic bias**: Slides of a presentation by a librarian at the University of Massachusetts Amherst (and active Wikipedian) concern "Systemic Bias in Wikipedia: What It Looks Like, and How to Deal with It".[14]
- **Few users who edit Middle East/North Africa articles are from the region**: A brief conference paper titled "The vocal minority: Local self-representation and coediting on Wikipedia in the Middle East and North Africa"[15] (presented in a slightly different form at a Workshop at 2012 ACM Web Science Conference [16] in June) analyzed the talk pages of English Wikipedia users who had edited articles geotagged in that region (MENA) "to assess the self-declared locational affiliations of the authors (i.e. where they live, work or were born)" and found that "there exists few authors claiming to be from the MENA region, except for Israel, Iran and to a much lesser extent Egypt."
- **Article Feedback tool as means of "peripheral participation"**: A paper to be presented at CSCW '13[17] describes the main findings from the early tests of the Article Feedback v5 on the English Wikipeida, from the lens of legitimate peripheral participation theory. The study reviews the costs and benefits of expanding reader contributions to Wikipedia, using both quantitative and qualitative methods. The results, according to the authors (members of the Wikimedia Foundation team working on the tool), indicate that peripheral contributors add value to the encyclopedia as long as the cost of identifying low quality contributions remains low.
- **Dynamics of read and edit rates on Wikipedia**: The ECCS'12 Conference on Complex Systems saw the presentation of a paper titled "From Time Series to Co-Evolving Functional Networks: Dynamics of the Complex System 'Wikipedia'"[18], reporting on research about the "access-rate time series and edit-interval time series" of articles on the English Wikipedia, and about " three organizational and dynamical networks ...: (i) the network of direct links between Wikipedia articles, (ii) the usage network as determined from cross-correlations between access-rate time series of many pairs of articles, and (iii) the edit network as determined from co-incident edit events. The major goal is to find correlations between components of these three networks that characterize the dynamics of information spread in the complex system".
- **Wikipedia articles compared to open source software projects**: A paper titled "Similarities, challenges and opportunities of Wikipedia content and open source projects"[19] argues that "the evolution of Wikipedia pages and the OSS projects share some commonalities in terms of their evolutionary patterns; in particular, it was found that a predefined, cubic model could be used to explain several of the similarities in 'abandoned' or 'completed'

projects and Wikipedia pages."

References

[1] Schroeder, A., Wagner, C. (2012). Governance of open content creation: A conceptualization and analysis of control and guiding mechanisms in the open content domain. *Journal of the American Society for Information Science and Technology* 63(10):1947–59 **DOI** (http://dx.doi.org/10.1002/asi.22657) Closed access

[2] Hirth, M., Lehrieder, F., Oberste-Vorth, S., Hossfeld, T., Phuoc T.-G. (2012). Wikipedia and its network of authors from a social network perspective. *2012 Fourth International Conference on Communications and Electronics* (ICCE) **DOI** (http://dx.doi.org/10.1109/CCE.2012.6315882) Closed access

[3] Douglas S. Schultz, Virginia M. Brabender: More Challenges Since Wikipedia: The Effects of Exposure to Internet Information About the Rorschach on Selected Comprehensive System Variables **PDF** (http://www.tandfonline.com/doi/abs/10.1080/00223891.2012.725438) Closed access"

[4] Randall,W. A. E. (2010). Rorschach reliability with exposure to Internet-based images and information (Unpublished doctoral dissertation). Massachusetts Professional School of Psychology, Boston, MA.

[5] Crowston, K., Jullien, N., Ortega, F. (in press) Is Wikipedia Inefficient? Modelling Effort and Participation in Wikipedia. *HICSS '13*, **PDF** (http://crowston.syr.edu/sites/crowston.syr.edu/files/hicss2013_CrowstonJullienOrtegawork_revised.pdf) Open access

[6] Colón-Aguirre, M. and Fleming-May, R. (in press). "'You just type in what you are looking for': Undergraduates' use of library resources vs. Wikipedia". *The Journal of Academic Librarianship.* (http://www.sciencedirect.com/science/article/pii/S0099133312001462)

[7] Wu, K.,Zhu, Q.,Vassileva, J.,Zhao, Y. (2012) Does conflict matter in the success of mass collaboration? Investigating antecedents and consequence of conflict in Wikipedia. *Chinese Journal of Library and Information Science*, 2012, 5(1):34-50 **PDF** (http://ir.las.ac.cn/handle/12502/5323) Open access

[8] Sérasset, G. (2012) Dbnary: Wiktionary as a LMF based Multilingual RDF network. *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12) **PDF** (http://www.lrec-conf.org/proceedings/lrec2012/pdf/387_Paper.pdf) Open access

[9] Capiluppi, Andrea; Duarte Pimentel, Ana Claudia; Boldyreff, Cornelia (04). " Patterns of creation and usage of Wikipedia content (http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6320537)". *Web Systems Evolution (WSE), 2012 14th IEEE International Symposium on*: 85-89. : 10.1109/WSE.2012.6320537 (http://dx.doi.org/10.1109/WSE.2012.6320537). Retrieved on 29 October 2012. Open access

[10] http://stats.grok.se/

[11] Gavin Baily: Wikisym 2012 Report (https://uk.wikimedia.org/wiki/Wikisym_2012_Report), Wikimedia UK wiki, October 2012

[12] Bahar Sateli: Wiki-NLP Integration at the WikiSym'12 Conference (http://www.semanticsoftware.info/blog/wiki-nlp-integration-wikisym12-conference), semanticsoftware.info blog, 2012-10-09

[13] Michael E. Kummer, Marianne Saam, Iassen Halatchliyski, and George Giorgidze: Centrality and Content Creation in Networks – The Case of German Wikipedia. ZEW Discussion Paper No. 12-053 **PDF** (ftp://ftp.zew.de/pub/zew-docs/dp/dp12053.pdf)

[14] Laura Quilter. "Systemic Bias in Wikipedia: What It Looks Like, and How to Deal with It". Open Access Week 2012, University of Massachusetts Amherst. Amherst, MA. Oct. 2012. **HTML** (http://works.bepress.com/laura_quilter/20) Open access

[15] Bernie Hogan, Mark Graham, Ahmed Medhat Mohamed: The vocal minority: Local self-representation and coediting on Wikipedia in the Middle East and North Africa. #Influence12 – Symposium & Workshop on Measuring Influence on Social Media, September 28–29, 2012 **PDF** (http://socialmedialab.ca/influence12/submissions/influence12_submission_37.pdf) Open access

[16] http://people.lis.illinois.edu/~jdiesner/calls/WON_2012.html

[17] Aaron Halfaker, Oliver Keyes, Dario Taraborelli: Making peripheral participation legitimate: Reader engagement experiments in Wikipedia. **PDF** (http://www-users.cs.umn.edu/~halfak/publications/Making_Peripheral_Participation_Legitimate/halfaker13making-preprint.pdf) Open access

[18] Mirko Kämpf, Jan. W. Kantelhardt, Lev Muchink: From Time Series to Co-Evolving Functional Networks: Dynamics of the Complex System 'Wikipedia' *ECCS'12 Conference on Complex Systems* **PDF** (http://85.214.43.8/ECCS2012-paper-v6.pdf) Open access

[19] Andrea Capiluppi: Similarities, challenges and opportunities of Wikipedia content and open source projects. Journal of Software: Evolution and Process **DOI** (http://dx.doi.org/10.1002/smr.1570) Closed access

Issue 2(11): November 2012

Movie success predictions, readability, credentials and authority, geographical comparisons

With contributions by: Piotrus, Benjamin Mako Hill, Tbayer, DarTar, Adler.fa, Hfordsa, Drdee

Early prediction of movie box-office revenues with Wikipedia data

An open-access preprint[1] has announced the results from a study attempting to predict early box-office revenues from Wikipedia traffic and activity data. The authors – a team of computational social scientists from Budapest University of Technology and Economics, Aalto University and the Central European University – submit that behavioral patterns on Wikipedia can be used for accurate forecasting, matching and in some cases outperforming the use of social media data for predictive modeling. The results, based on a corpus of 312 English Wikipedia articles on movies released in 2010, indicate that the joint editing activity and traffic measures on Wikipedia are strong predictors of box-office revenue for highly successful movies.

The authors contrast their *early prediction* approach with more popular *real-time prediction/monitoring* methods, and suggest that movie popularity can be accurately predicted well in advance, up to a month before the release. The study received broad press coverage and was featured in *The Guardian*, the *MIT Technology Review* and the *Hollywood Reporter* among others. The authors observe that their approach, being "free of any language based analysis, e.g., sentiment analysis, could be easily generalized to non-English speaking movie markets or even other kinds of products". The dataset used for this study, including the financial and Wikipedia activity data is available among the supplementary materials of the paper.

Readability of the English Wikipedia, Simple Wikipedia, and Britannica compared

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$

The automated readability index, one of the readability metrics used in the study[2]

A study[2] by researchers at Kyoto University presents a detailed assessment of the readability of the English Wikipedia against Encyclopedia Britannica and the Simple English Wikipedia using a series of readability metrics and finds that Wikipedia "seems to lag behind the other encyclopedias in terms of readability and comprehensibility of its content". The paper, presented at *CIKM'12*, uses a variety of metrics spanning syntactical readability indices (such as Flesch reading ease, the automated readability index and the Coleman–Liau index) as well as metrics based on word popularity (including the Dale–Chall readability formula and word frequency indices derived from Google News or the American National Corpus).

The authors prepared a corpus of matching articles for the purpose of comparison between the English and Simple English Wikipedia. It should be noted that the authors didn't perform a random selection of articles, but selected a sample based on the existence of a corresponding article in Simple Wikipedia. The findings of the first analysis indicate that Simple Wikipedia consistently outperforms the English Wikipedia on all readability metrics. Wikipedia also appears to contain on average more proper nouns than Britannica – which, the authors speculate, may be due to specific editorial policies. The second section of the paper measures readability for 500 articles for each one of eight topic categories selected from DBpedia (biology, chemistry, computing, economics, history, literature, mathematics, and philosophy).

The comparison indicates that articles in the *computing* category are the most readable by syntactical and familiarity measures. *Biology* and *chemistry*, on the other hand, seem to include the most difficult articles. The final section reviews the readability of Britannica articles, in particular comparing the readability of articles in the "introductory" class with that of Simple Wikipedia articles and the readability of "encyclopedia" class articles with that of

Wikipedia articles. The findings indicate that Britannica outperforms Wikipedia in readability overall, while introductory articles outperform Simple Wikipedia articles. It should be noted that the comparisons were not performed on matched pairs and that the authors do not specify what criteria were used to sample articles from Britannica.

A paper whose preprint was previously covered in this research report, and now published as a full research article in PLOS One,[3] found that the Simple English Wikipedia has a higher degree of complexity than the corpus of Charles Dickens' books when measured via the Gunning fog index, but is less complex than the British National Corpus, "which is a reasonable approximation to what we would want to think of as 'English in general'". See also the September issue of this research report for a summary of a third readability study which had applied the standard Flesch Reading Ease test to the English and Simple English Wikipedias.

Wikipedia favors established views and scientifically backed knowledge

An article appearing in *Information, Communication & Society*[4] studies the discussion pages of English and German September 11 attacks articles, contributing to the ongoing debates on collaborative knowledge creation in the wiki Web 2.0 context, participation of experts and amateurs on Wikipedia, and, indirectly, reliability of Wikipedia. The article's research question, coming from the sociology of knowledge and social constructivism perspectives, asks to what degree Wikipedia's "anyone can edit" policy democratizes the production of knowledge, removing it from traditional hierarchies "between experts and lay participants". The term *democratization* here is used in the context of such theoretical concepts as wisdom of crowds, participatory culture, produsage and (more critically) the notions of cult of the amateur or digital Maoism. All of these refer to the fact that Wikipedia's editors are more often amateurs ("lay participants") than professionally recognized experts.

Using the grounded theory approach, the study focuses not on editors, but on their arguments. It finds that due to community-upheld Wikipedia policies such as Wikipedia:Reliable sources, dissenting opinions ("traditionally marginalized types of knowledge") such as various conspiracy theories are still marginalized or straight-out excluded; according to the author, this "did not lead to a 'democratization' of knowledge production, but rather re-enacted established hierarchies". The finding should be taken in a certain context; as the author notes, the article was written by amateurs ("lay participants"), who however decided to reproduce traditional knowledge hierarchies, relegating various conspiracy theories and similar points not backed up to reliable sources to obscurity on Wikipedia. The author also concludes that Wikipedia, like other encyclopedias, is prone to a "scientism bias", i.e. treating scientifically backed knowledge as "better" than knowledge coming from alternative outlets. This despite the "anyone can edit" motto of Wikipedia, the author finds support for the argument that Wikipedia puts more stress on article quality than democratic participation, or in the words of the author: "Although laypeople apparently play a significant part in the text production, this does not mean that they favor lay knowledge. On the contrary, it is clearly elite knowledge of well-established authorities which is finally included in the article, whereas alternative interpretations are harshly excluded or at least marginalized."

Side-note: This reviewer found the author's use of a Firefox add-on Wired-Maker [5] for content analysis rather ingenious, and applauds the mentioning of such a practical methodological tip in their paper.

Trust, authority and credentials on Wikipedia: The case of the Essjay controversy

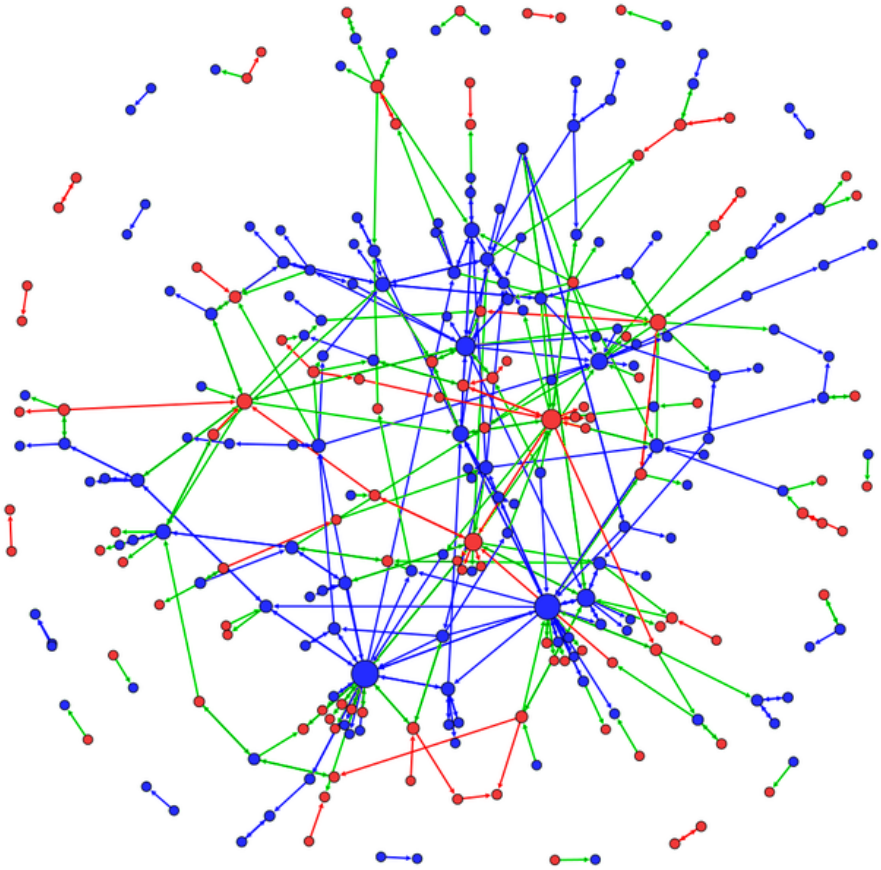
At the Academy of Management conference in Boston, Dariusz Jemielniak presented a paper on *Trust, Control, and Formalization in Open-Collaboration Communities: A Qualitative Study of Wikipedia* [6]. It is built around a detailed description and interpretation of the Essjay controversy on the English Wikipedia in 2007 about the use of inaccurate credentials by active Wikipedian and administrator Essjay. The paper is framed in terms of the literature from organization theory on trust and control. Jemielniak argues that organization theory suggests that organizations must either be able or willing to trust participants or must rely on control systems which essentially obviate the need for trust. Using ethnographic data from Wikipedia, Jemielniak suggests that Wikipedia — and, perhaps, a series of

similar computer-mediated "open-collaboration communities" — instead rely on a series of procedures and "legalistic remedies" which provide a previously untheorized alternative to traditional control systems used in organizations.

The working paper is the first in what Jemielniak suggests will be a series of papers based on a long-term participatory ethnographic study: over the past five years, Jemielniak has edited Wikipedia almost daily and is a steward on Wikimedia projects (as well as the chair of the Wikimedia movement's newly established Funds Dissemination Committee, and recently announced the committee's recommendations on funding requests by various Wikimedia organizations totaling US$10.4M). Jemielniak uses his own experience as well as detailed on-wiki records from conversations surrounding the Essjay affair to walk through the controversy and its implications in depth. He discusses how Wikipedians construct authority and initially reacted with indifference to the revelation that Essjay had used fake credentials, how this changed when new information about Essjay's use of his credentials came to light, how a series of proposals to prevent or respond to such issues in the future were raised, and how the community essentially decided to keep the *status quo*.

The paper paints a detailed, nuanced, and deeply informed portrait of Wikipedians' responses to the controversy and the ways in which trust and its relationships to authority and credentials are navigated in the project. The author suggests that the creation of rules and legalistic procedures allowed Wikipedians to walk the line between rejecting descriptions of authority *per se* while minimizing the effects of inaccurate descriptions of authority by suggesting that editors on Wikipedia should rely much more heavily on users' experience and on the degree to which particular contributions conform to Wikipedia's content guidelines.

A working paper by the same author, presented at the annual meeting of the Society for Applied Anthropology[7] gives an overview of Wikipedia's culture by reviewing the role of its norms, guidelines and policies.



Network of users communicating on Wikipedia article talk pages (Neff et al., p.22).[8] Edges connecting two Democrats are colored blue, edges connecting two Republicans in red, and edges representing inter-party dialogue are shown in green.

Briefly

- **Being Wikipedian is more important than the political affiliation**: In a recent preprint[8], titled "Jointly they edit: examining the impact of community identification on political interaction in Wikipedia", researchers have studied the political identity of Wikipedia editors and investigated the effect of the social identity on their editorial activity patterns. The paper starts with a long and comprehensive review of the concept of social identity and generalizes it to online social identity. Based on an analysis of a sample of 1390 editors with known political affiliation – either US Democrat or Republican – they conclude that although the social identity of editors is strongly reflected in their editorial interests – that is, the topics on which they are more active – but that being "Wikipedian" dominates political affiliation when it comes to user pages. In contrast with other social media e.g., the blogosphere, where cross-party interactions are very much underrepresented, it appears that Wikipedian dialogues between editors from opposing parties are relatively profound and notable. On the day before the US presidential election, the paper's results were highlighted on the Wikimedia blog under the headline "In divisive times, Wikipedia brings political opponents together [9]".
- **Eye-tracking study: Readers look at TOC first, then infobox**: A conference paper titled "Looking for genre: the use of structural features during search tasks with Wikipedia"[10] described the results of an eye tracking study, where readers looking for information in a Wikipedia article tended to look first at the table of contents, then at the article's infobox. Also, readers frequently "skim and scroll" long articles.
- **Edit categories in featured and non-featured articles**: This article focuses on some differences between featured and non-featured article. Unsurprisingly, the authors main finding is that the featured articles are more stable after promotion; the interesting contribution of the authors lies more in their detailed methodology, and categorization of various types of edits.[11]
- **How the TV schedule influences Wikipedia pageviews**: In Germany, several recent consumer studies have found evidence for a rise of what has been called "second screen": The parallel use of TV and the Internet. To find a partial answer to the question whether this use is unrelated (e.g. checking emails while the TV is running in the background) or integrates both media, a blogger turned to pageview numbers for the German Wikipedia[12]. From a still unsystematic analysis, he draws two conclusions: "First, the use of Wikipedia is markedly influenced by the TV schedule. On Saturday evenings in particular, but sometimes also during weekdays, the most viewed Wikipedia entries contain many articles related to the currently showing TV program. Secondly, these articles are primarily viewed while the corresponding show is running on TV." The author also announced a Perl script to convert the raw pageview data provided by the Wikimedia Foundation [14] into a MySQL database, demonstrated in a live list of the 50 most viewed articles [13] of the German Wikipedia.
- **A truthfulness verification system based on Wikipedia**: Yang Liu's master's thesis (paywalled)[14]discusses the development of WT-verifier, a "a truthfulness verification system based on Wikipedia" that uses information on Wikipedia, rather than general web searches to perform fact checking. Liu finds that Wikipedia "has high reliability of page contents, due to strict rules for page editing and a strong self-fixing mechanism" and adapts T-verifier, an existing system based on Yahoo! searches, applying it to information on Wikipedia. Liu develops what he calls a "truthfulness aware snippet generation algorithm" and finds that the new approach "significantly increases the precision and recall compared to the original T-verifier approach."
- **Characterizing Wikipedia traffic**: A paper presented at the *7th International Conference on Internet and Web Applications and Services*[15] gives a breakdown of Wikipedia traffic for 2009 to the 10 largest wikis with a particular focus on content-type. The authors give a high-level overview of Wikipedia traffic but they do not take the opportunity to dive deeper into the data. The current analysis from the paper can also be found on the Wikimedia Report Card [16] and Wikimedia Statistics page [17]. Suggestions for future research include the following: an in-depth analysis of the temporal dynamics of editing behavior. For example, do we see higher editor activity during holidays? An in-depth analysis of the multi-media files/Wikimedia Commons project. Are there differences between wiki projects regarding the use of Commons image files?

- **One-year article ratings dump released**: The Wikimedia Foundation announced the release of the complete, anonymous data dump of 11M article ratings collected over 1 year (July 2011 – July 2012) from the English Wikipedia via the article feedback tool (v4).[18] The dataset is released under a CC0 license.
- **Measuring countries' visibility on Wikipedia:** On his "Zero Geography" blog, researcher Mark Graham began a series of posts[19][20] comparing the "geography of views" for different countries on Wikipedia: "we constructed a list of every single article about a place (towns, monuments, historical events, rivers, buildings etc.) in the top 42 Wikipedia language versions, and then queried the number of views that each of those articles received over a two-year period (2009-2011)." (This is part of ongoing research into geographical aspects of the information on Wikipedia by Graham's team at the Oxford Internet Institute, and will be featured in an upcoming paper.) Content about US locations received the most views across languages, followed by the UK and Germany. Graham observed that the top 10 list by pageviews shows a lot of similarity to the top 10 lists of countries by number of articles, and by number of edits originating from that country, but noted that "the UK [being] Europe's most visible country ... is quite interesting because it isn't the country in Europe that uses Wikipedia the most (Germany does)", conjecturing that this might have to do with language differences.
- **Ratio of African Wikipedia readers rising, but still low**: Erik Zachte, data analyst at the Wikimedia Foundation, blogged an update about "Wikipedia page reads, breakdown by region"[21], observing among other things that "Africa still has a long way to go to gain equal access to internet: with about 15% of the worlds population [22], 1.4 % of Wikipedia page views is low, but still one and a half as much as 3 years ago."

References

[1] Mestyán, M., Yasseri, T., & Kertész, J. (2012). Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data. *ArXiV*. **PDF** (http://arxiv.org/abs/1211.0970) Open access

[2] Jatowt, A., & Tanaka, K. (2012). Is Wikipedia Too Difficult? Comparative Analysis of Readability of Wikipedia , Simple Wikipedia and Britannica. *CIKM'12*, pp. 2607–2610. **PDF** (http://www.dl.kuis.kyoto-u.ac.jp/~adam/cikm12a.pdf) • **DOI** (http://dx.doi.org/10.1145/2396761.2398703) Open access

[3] Yasseri, T., Kornai, A., & Kertész, J. (2012). A Practical Approach to Language Complexity: A Wikipedia Case Study. *PLoS ONE*, 7(11), e48386. **DOI** (http://dx.doi.org/10.1371/journal.pone.0048386) Open access

[4] König, R. (2012). Wikipedia. Between lay participation and elite knowledge representation. *Information, Communication & Society*. Advance online publication. **DOI** (http://dx.doi.org/10.1080/1369118X.2012.734319) Closed access

[5] https://addons.mozilla.org/en-US/firefox/addon/wired-marker/

[6] Jemielniak, D. (2012). Trust, Control, and Formalization in Open-Collaboration Communities: A Qualitative Study of Wikipedia. *Academy of Management 2012 Annual Meeting*. **PDF** (http://depot.ceon.pl/handle/123456789/308) Open access

[7] Jemielniak, D. (2012). Wikipedia: An effective anarchy. *Society for Applied Anthropology 2012 Annual Meeting* (SfAA 2012). **PDF** (https://depot.ceon.pl/handle/123456789/315) Open access

[8] Neff, J. G., Laniado, D., Kappler, K., Volkovich, Y., Aragón, P., & Kaltenbrunner, A. (2012). Jointly they edit: examining the impact of community identification on political interaction in Wikipedia. *ArXiV*, **PDF** (http://arxiv.org/abs/1210.6883) Open access

[9] https://blog.wikimedia.org/2012/11/05/in-divisive-times-wikipedia-brings-political-opponents-together/

[10] Clark, Malcolm; Ruthven, Ian; O'Brian Holt, Patrik and Song, Dawei (2012). Looking for genre: the use of structural features during search tasks with Wikipedia. *Fourth Information Interaction in Context Conference* (IIiX 2012). (http://dx.doi.org/10.1145/2362724.2362751'"DOI"') • **PDF** (http://oro.open.ac.uk/34649/1/04-iiix2012_submission_26.pdf) Open access

[11] Daxenberger, J., & Gurevych, I. (2012). A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. *Proceedings of the 24th International Conference on Computational Linguistics* (COLING 2012). **PDF** (http://www.ukp.tu-darmstadt.de/data/textual-revisions/wikipedia-edit-category-corpus/) Open access

[12] Rycak, M. (17 November, 2012) Wikipedia-Zugriffszahlen bestätigen Second-Screen-Trend. *martinrycak.de*. **HTML** (http://www.martinrycak.de/wikipedia-zugriffszahlen-bestatigen-second-screen-trend/) Open access

[13] http://martin.rycak.de/wikitrends/last.html

[14] Liu, Y. (2012). WT-verifier. Truthfulness verification of fact statements on Wikipedia (unpublished masters' thesis). State University of New York at Binghamton. **HTML** (http://gradworks.umi.com/15/16/1516645.html) Closed access

[15] Reinoso, A. J., Muñoz-Mansilla, R., Herraiz, I., & Ortega, F. (2012). Characterization of the Wikipedia Traffic. *Seventh International Conference on Internet and Web Applications and Services* (ICIW 2012), pp. 156–162. **PDF** (http://www.thinkmind.org/index.php?view=article&articleid=iciw_2012_5_50_20194) Open access

[16] http://reportcard.wmflabs.org

[17] http://stats.wikimedia.org

[18] Taraborelli, D. (2012) Wikipedia article ratings. *The Data Hub* **TSV** (http://datahub.io/dataset/wikipedia-article-ratings) Open access

[19] Graham, M. (5 November 2012). Virtuous Visible Circles: mapping views to place-based Wikipedia articles. *Zero Geography*. **HTML** (http://www.zerogeography.net/2012/11/virtuous-visible-circles-mapping-views.html) Open access

[20] Graham, M. (11 November 2012). The most visible country in Europe (on Wikipedia) is... *Zero Geography*. **HTML** (http://www.zerogeography.net/2012/11/the-most-visible-country-in-europe-on.html) Open access

[21] Zachte, E. (15 November 2012) Wikipedia page reads, breakdown by region. *Infodisiac*. **HTML** (http://infodisiac.com/blog/2012/11/wikipedia-page-reads-breakdown-by-region/) Open access

[22] http://en.wikipedia.org/wiki/World_population

Issue 2(12): December 2012

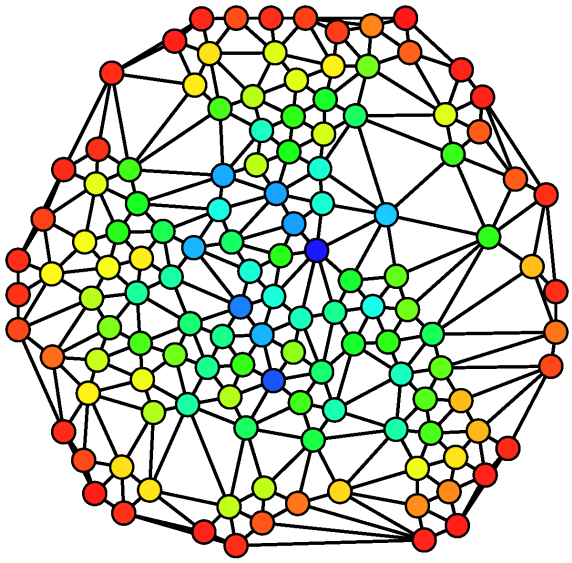
Wikipedia and Sandy Hook; SOPA blackout reexamined

With contributions by: Daniel Mietchen, Piotrus, Junkie.dolphin, Taha Yasseri, Benjamin Mako Hill, Aaron Shaw, Tbayer, DarTar and Ragesoss

How Wikipedia deals with a mass shooting

Northeastern University researcher Brian Keegan analyzed the gathering of hundreds of Wikipedians to cover the Sandy Hook Elementary School shooting in the immediate aftermath of the tragedy. The findings are reported in a detailed blog post that was later republished by the Nieman Journalism Lab.[1] Keegan observes that the Sandy Hook shooting article reached a length of 50Kb within 24 hours of its creation, making it the fastest growing article by length in the first day among recent articles covering mass shootings on the English-language Wikipedia. The analysis compares the Sandy Hook page with six similar articles from a list of 43 articles on shooting sprees in the US since 2007. Among the analyses described in the study, of particular interest is the dynamics of dedicated vs occasional contributors as the article reaches maturity: while in the first few hours contributions are evenly distributed with a majority of single-edit editors, after hour 3 or 4 a number of dedicated editors show up and "begin to take a vested interest in the article, which is manifest in the rapid centralization of the article". A plot of inter-edit time also shows the sustained frequency of revisions that these articles display days after their creation, with Sandy Hook averaging at about 1 edit/minute around 24 hours since its first revision. The notebook and social network data produced by the author for the analysis are available on his website [2]. The Nieman Journalism Lab previously covered the role that Wikipedia is playing as a platform for collaborative journalism, and why its format outperforms Wikinews with an interview of Andrew Lih published in 2010.[3] The early revision history of the Sandy Hook shooting article was also covered in a blog post by Oxford Internet Institute fellow Taha Yasseri, however with a focus on the coverage in different Wikipedia language editions.[4]

Network positions and contributions to online public goods: the case of the Chinese Wikipedia



A graph with nodes color-coded by betweenness centrality (from red=0 to blue=max).

In a forthcoming paper in the *Journal of Management Information Systems* (presented earlier at HICSS '12[5]), Xiaoquan (Michael) Zhang [6] and Chong (Alex) Wang [7] use a natural experiment to demonstrate that changes to the position of individuals within the editor network of a wiki modify their editing behavior. The data for this study came from the Chinese Wikipedia. In October 2005, the Chinese government suddenly blocked access to the Chinese Wikipedia from mainland China, creating an unanticipated decline in the editor population. As a result, the remaining editors found themselves in a new network structure and, the authors claim, any changes in editor behavior that ensued are likely effects of this discontinuous "shock" to the network. The paper defines each editor as a node (vertex) in the network and a tie (edge) between two editors is created whenever the editors edit the same page in the wiki. They then examine how changes to three aspects of individual editors' relative connectedness (centrality) to other editors within the network altered their subsequent patterns of contribution.

The main finding is that changes in the three kinds of editors' connectedness within the network result in differential changes to their editing behavior. First, an increase in the number of direct connections between one editor and the rest of the network (degree centrality) resulted in fewer edits by that editor, and more work on articles they created. Second, an increase in the overall proximity of an editor to the other members of the network (closeness centrality) resulted in fewer edits and less work on articles they created. Third, an increase in the extent to which an editor connected otherwise isolated groups in the network (betweenness centrality) resulted in more edits and more work by that editor on articles they created. Overall, these results imply that alterations to the network structure of a wiki can change both the quantity and quality of editor contributions. The researchers argue that their findings confirm the predictions of both network game theory and role theory; and that future research should try to analyze the character of the network ties created within platforms for large-scale online collaboration, to better understand how changes to network structure may alter collaborative practices and public goods creation.

Quality of pharmaceutical articles in the Spanish Wikipedia

Ibuprofen, one of the World Health Organisation's "essential drugs", a topic covered in detail by the Spanish-language Wikipedia.

In an online early version of an upcoming article in *Atención Primaria*,[8] researchers at the Miguel Hernández University of Elche and the University of Alicante have benchmarked articles on pharmaceutical drugs in the Spanish Wikipedia against information available in a pharmaceutical database, *Vademécum*.[9] A subset of the Vademécum corpus of 3,595 drugs was created using simple random sampling without replacement, consisting of 386 drugs. Of these, 171 (44%) had entries on the Spanish Wikipedia, which were then scrutinized along several dimensions in May 2012. Usage of the drug was correctly indicated in 155 (91%) of these articles, dosage in 26 (15%), and side-effects in 64 (37%), with only 15 articles (9%) scoring well in all of these dimensions. The researchers conclude that, while Wikipedia has a high potential to help with the dissemination of pharmaceutical knowledge, the Spanish-language edition does not currently live up to this potential. As a possible solution, they suggest the pharmaceutical community more actively participate in editing Wikipedia. The list of the drugs involved has not been made public, since a similar study is currently underway whose results may be distorted by targeted intervention. The authors have signalled to this research report their intention to make the list available after this second study is complete.

Wikipedia editing patterns are consistent with a non-finite state model of computation

A paper posted to ArXiv[10] by SFI's Omidyar fellow Simon DeDeo [11] presents evidence for non-finite state computation in a human social system using data from Wikipedia edit histories. Finite state-systems are the basis for the study of formal languages in computer science and linguistics, and many real-world complex phenomena in biology and the social sciences are also studied empirically by assuming the existence of underlying finite-state processes, for the analysis of which powerful probabilistic methods have been devised. However, the question of whether the description of a system truly entails a finite or a non-finite, unbounded number of states, is an open one. This is significant from a functionalist point of view: can we classify a system by its computational properties, and can these properties help us better understand how the system works regardless of its material details?

The paper's contribution lies in its proof of a probabilistic generalization of the pumping lemma, a device used in theoretical computer science as a necessary condition for a language to be described by only a finite number of states. The lemma is applied to the edit histories of a number of the most frequently edited articles in the English Wikipedia, after being properly transformed into coarse-grain sequences of "cooperative" or "non-cooperative (reversion) edits (reverts being identified by means of their SHA1 field). A Bayesian argument is applied to show that the lemma cannot hold for a majority of sequences, thus showing that Wikipedia's collaborative editing system as a whole cannot be described by any aggregation of finite-state systems. The author discusses the implications of this finding for a more grounded study of Wikipedia's editing model, and for the identification of detailed models of computation for other social and biological systems.

Wikipedia as our collective memory

Michela Ferron, a member of the SoNet (Social Networking) research group [12] at the Bruno Kessler Foundation in Trento, Italy submitted her PhD thesis[13] in December 2012. She examined the idea of viewing Wikipedia as a venue for collective memory and the language indicators of the dynamic process of memory formation in response to "traumatic" events. Parts of the thesis have already been published in journals and conference proceedings, such as WikiSym 2011 and 2012 (cf. presentation slides [14]).

A protester on Tahrir Square during the 2011 Egyptian revolution.

A full chapter is dedicated to the background on the concept of collective memory and its appearance in the digital world. The thesis continues with an analysis of "anniversary edits", showing a significant increase in editorial activities on articles related to traumatic events during the anniversary period compared to a large random sample of "other" articles. More detailed linguistic indicators are introduced in the next chapter. It is statistically shown that the terms related to affective processes, negative emotions, and cognitive and social processes occur more often in articles on traumatic events; "Specifically, the relative number of words expressing anxiety (e.g., “worried”), anger (e.g., “hate”) and sadness (e.g., “cry”) was significantly higher in articles about traumatic events".

In the next step, Ferron tried to distinguish between human-made and natural disasters. It has been observed that "human-made traumatic events were characterized by language referring to anger and anxiety, while the collective representation of natural disasters expressed more sadness". Finally, a detailed case study of the talk pages of articles on the 7 July 2005 London bombings and the 2011 Egyptian revolution was carried out, and language indicators, especially those related to emotions, were investigated in a dynamic framework and compared for both examples.

SOPA blackout decision analyzed

A *First Monday* article[15] reviews several aspects of the Wikipedia participation in the 18 January 2012 protests against SOPA and PIPA legislation in the US. The paper focuses on the question of legitimacy, looking at how the Wikipedia community arrived at the decision to participate in those protests.

The English Wikipedia landing page, symbolically its only page during the blackout on January 18, 2012

The paper provides an interesting discussion of legitimacy in Wikipedia's governance, and discusses the legitimacy of the decision to participate in the protests. The author notes that the initiative was given a major boost by Jimmy Wales' charismatic authority, as Wales posted a straw poll about the issue on his talk page on December 10, 2011, as while the issue was discussed by the community beforehand (for example, in mid-November at the Village Pump), those discussions attracted much less attention. It is hard to say whether the protest would have happened without Jimbo's push for more discussion, as it veers towards "what if" territory; as things happened, it is true that Jimbo's actions began a landslide that led to the protests. However, this reviewer is more puzzled at the claim made in the introduction to the article that the discussion involved a "massive involvement of the Wikimedia Foundation

staff". While several WMF staffers were active in the discussions in their official capacity, and while the WMF did issue some official statements about the ongoing discussion, the paper certainly does not provide any evidence to justify the word "massive".

The paper subsequently notes that the WMF focused on providing information and gently steering the discussion, without any coercion; this hardly justifies the claim of "massive involvement". At the very least, a clear explanation is necessary of precisely how many WMF staffers participated in the discussion before such a grandiose adjective as "massive" is used. It is true that the WMF staffers helped push the discussion forward, but this reviewer believes that the paper does not sufficiently justify the stress it puts on their participation, and thus may overestimate their influence.

The third part of the paper discusses how the arguments about legitimacy or the lack of it framed the subsequent discourse of the voters. The author notes that after initial period of discussing SOPA itself, the discussion of whether it was legitimate or not for Wikipedia to become involved in the protest took over, with a major justification for it emerging in the form of an argument that it was legitimate for Wikipedia to protest against SOPA as SOPA threatened Wikipedia itself. While this is an interesting claim, unfortunately, other than citing one single comment, no other qualitative or quantitative data are provided; nor is the methodology discussed. We are not told how many individuals voted, how many commented on legitimacy or illegitimacy, how many felt that Wikipedia is threatened; we do not know how the author classified comments supporting any of the viewpoints, or the shifts in the discussion ... this list could unfortunately go on. In one specific example drawn from the conclusion, the author writes that "The main factor that shaped the multi-phased process was the will to have the community accept the final decision as legitimate, and avoid backlash. This factor especially influenced those who are suspected of relying on traditional means of legitimacy such as charisma or professionalism." At the same time, we are provided with no number, no percentage, and certainly no correlation to back up this claim. Without a clear methodology or distinct data it is hard to verify the author's claims and conclusions.

The introduction also notes that "the mass effort of planning an effective political action was not something “anyone [could] edit”" and "the debate preceding the blackout did not follow Wikipedia’s open and anarchic decision-making system"; unfortunately this reviewer finds no justification for those rather strong claims anywhere else in the article.

Overall, this is an interesting paper about legitimacy in Wikipedia, but it seems to overreach when it tries to draw conclusions from the data that is simply not presented to the reader. It suffers from a failure to explain the research's methodology, making verification of the claims made very hard. Due to the lack of hard data, most conclusions are unfortunately rendered dubious, and the paper has a tendency to make strong claims that are not backed up by data or even developed later on.

Bots and collective intelligence explored in dissertation

In his Communication and Society PhD dissertation,[17] Randall M. Livingstone of the University of Oregon explores the relationship between the social and technical structures of Wikipedia, with a particular focus on bots and bot operators. After a fairly broad literature review (which summarizes the basic approaches to Wikipedia studies from new media theory, social network analysis, science and technology studies, and political economy), Livingstone gives a concise history of the technical development of Wikipedia, from UseModWiki to MediaWiki, and from a single server to hundreds.

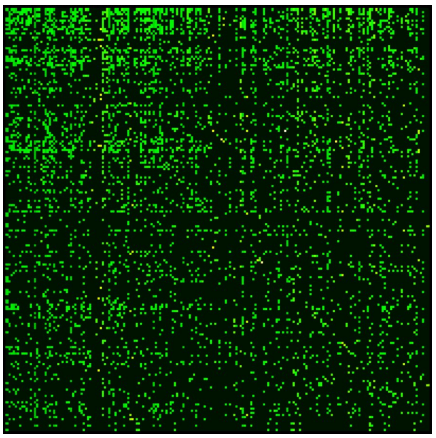
Rats (blue trace) interacting with a rat-sized robot (red) controlled by a human who in turn perceives the rat's movements through those of a human-sized avatar in a virtual reality environment.[16] The video was uploaded to Wikimedia Commons by the Open Access Media Importer Bot.

The most interesting chapters for Wikipedians will be V – Wikipedia as a Sociotechnical System – and VI – Wikipedia as Collective Intelligence. Chapter 5 looks at the ways the editing community and the evolution of software (both MediaWiki and the semi-automated tools and bots that interact with editors and articles) "construct" each other. Based on 45 interviews with bot operators and WMF staff, this chapter gives an interesting and varied picture of how Wikipedia works as a sociotechnical system. It will in part be a familiar account to the more tech-minded Wikipedians, but offers an accessible overview of bots and their place in the ecosystem to editors who normally steer clear of bots and software development. Chapter 6 looks at theories of intelligence and the concept of collective intelligence, arguing that Wikipedia exhibits (at least to some extent) the key traits of stigmergy, distributed cognition, and emergence.

Briefly

- **"History's most influential people" according to Wikipedia**: While more in the realm of popular science, *Wired UK*, among others, published[18] an infographic attributed to César Hidalgo, head of the MIT Media Lab's Macro Connections group, visualizing "History's most influential people". Unfortunately, beyond noting that rankings "are based on parameters such as the number of language editions in which that person has a page, and the number of people known to speak those languages" the small article does not provide any methodology, nor does it provide much discussion. Until a more extensive description is released, the current graph, while pretty, is little more than a trivia piece.
- **Teachers say 75% of teens use Wikipedia (or online encyclopedias) for research assignments**: In a Pew Research survey among more than 2000 US middle and high school teachers[19] 75% said that their teenage students use "Wikipedia or other online encyclopedia" in research assignments, making online encyclopedias the second most popular source for students behind search engines such as Google. This number [20] was lower (68%) "among teachers of the lowest income students (those living below the poverty line)" and higher (80%) for those teaching "mostly upper and upper middle income" students, and it also varied by subject (between 69% for teachers of English and 82% for science teachers). The survey report cautions that the sample "skews towards 'cutting edge' educators who teach some of the most academically successful students in the country".

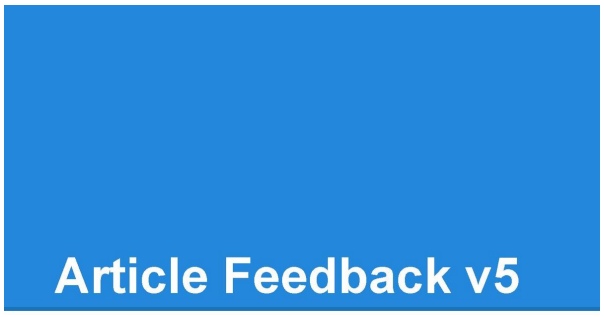
- **"Wikipedia communities" as eigenvectors of its Google matrix**: An ArXiv preprint[21] studies the "Spectral properties of Google matrix of Wikipedia and other networks". This Google matrix consists of entries for each pair of pages (for the English Wikipedia, including non-mainspace pages like portals), roughly speaking modelling the behavior of a surfer who goes from one page to any of those that it links to, with equal probability (or, with probability $1-\alpha$, jumps to a random page; the damping parameter α is set to around 0.85 in the Google search engine). The PageRank appears as the eigenvector of this matrix for the eigenvalue $\lambda = 1$. The paper studies the spectrum (eigenvalues) and eigenvectors apart from this special case, interpreting them as certain topic areas: "the eigenvectors of the Google matrix of Wikipedia clearly identify certain communities which are relatively weakly connected with the Wikipedia core when the modulus of corresponding eigenvalue is close to unity. For moderate values of $|\lambda|$ we still have well defined communities which however have stronger links with some popular articles (e.g. countries) that leads to a more rapid decay of such eigenmodes."



The Google matrix of Wikipedia entries, from an earlier paper by the same authors of this study.[21]

- **Serial singularities: developing of a network organization by organizing events**: In a paper published in the Schmalenbach Business Review,[22] Leonhard Dobusch and Gordon Müller-Seitz from the Freie Universität Berlin suggest that research on organized events has tended to treat those events as isolated and singular events. Using interviews and other data on Wikimania, chapter meetings, and local meet-ups over several years, the authors challenge this idea and show how many different events on different scales and scopes – each with a distinct character – can interact and reinforce each other to help drive the nature of a large distributed organization like Wikimedia.
- **The web mirrors value in the real world: comparing a firm's valuation with its web network position**: In a MIT Sloan Working Paper,[23] Qiaoyun Yun and Peter Gloor create a measure of US and Chinese firms "social network" position by looking at how those firms are linked to from a variety of web sources – prominently Wikipedia. They find a positive correlation between betweenness centrality of a firm in a social network constructed from links online and its innovation capability and financial performance. They find that Wikipedia only predicts a firm's performance in the US.
- **Teahouse analyzed**: Jonathan Morgan, Sarah Stierch, Siko Bouterse and Heather Walls, from the Wikimedia Foundation Teahouse team, report on the impact of the initiative on 1,098 new Wikipedia contributors who joined the Teahouse between February and October 2012, in a paper to be presented at *CSCW '13*.[24] The study reports that participants in the project "make more edits overall, and edit longer", "make more edits, to more articles" and "participate more in discussion spaces" compared to non-visitors. This paper is part of a research track entirely dedicated to Wikipedia Supported Collaborative Work [25], featuring three other studies.

- **Article feedback**: The Wikimedia Foundation published an update about the Article feedback tool on the English Wikipedia, providing statistics about the usage of the feature, and about the moderation activities for the feedback provided.[26]



Slides from the recently published Article Feedback research report.[26]

- **New review of *Good Faith Collaboration***: The reviewer locates[27] Joseph Reagle's 2010 book about Wikipedia (free online version [28]) as following in a wider context of research on Wikipedia: "The reliability of the encyclopaedia's content.. and quantitative analysis of large-scale public datasets formed the predominant approach in early empirical research on Wikipedia ... This was followed by a more social approach and the adopting of qualitative methods. In this switch to social norms and away from an ethnographic approach, Reagle's book is a main reference, particularly in terms of its cultural and historical specificity." Overall, the review finds that "The book is well documented, with an elaborative but accessible writing style, which is at times provocative. It results in a form of rich composition of eight pieces (chapters) of Wikipedia 'puzzle', even if some readers might miss a more explicit continuum linking the lines together. Finally, the book is a primary reference point for researchers aiming to study Wikipedia, especially for those unfamiliar with it."
- **Measuring the impact of Wikipedia for GLAM institutions**: Ed Baker [29], software developer at the Natural History Museum in London, has started a series of blog posts on "the impact and use of Wikipedia by organisations".[30] In the first post, he looked at how the scope of pages linking to the NHM's website fits with the overall scope of the institution when pages are ranked either by number of page views or by number of links to the NHM. The latter approach could help identify opportunities for a collaboration between GLAM institutions and the Wikimedia communities.

References

[1] Keegan, B. (2012). How does Wikipedia deal with a mass shooting? A frenzied start gives way to a few core editors. *Nieman Journalism Lab* **HTML** (http://www.niemanlab.org/2012/12/how-does-wikipedia-deal-with-a-mass-shooting-a-frenzied-start-gives-way-to-a-few-core-editors/) Open access

[2] http://www.brianckeegan.com/2012/12/sandy-hook-school-massacre/

[3] Seward, Z.M. (2012) Why Wikipedia beats Wikinews as a collaborative journalism project. *Nieman Journalism Lab* **HTML** (http://www.niemanlab.org/2010/02/why-wikipedia-beats-wikinews-as-a-collaborative-journalism-project/) Open access

[4] Yasseri, T. (2012) The coverage of a tragedy. *Stories for Sunday morning* (http://tahayasseri.wordpress.com/2012/12/17/the-coverage-of-a-tragedy/'"HTML"') Open access

[5] Wang, C. (Alex), & Zhang, X. (Michael). (2012). Network Centrality and Contributions to Online Public Good–The Case of Chinese Wikipedia. *2012 45th Hawaii International Conference on System Sciences* (pp. 4515–4524). IEEE. **DOI** (http://dx.doi.org/10.1109/HICSS.2012.444) Closed access

[6] http://mikezhang.com/

[7] https://sites.google.com/site/wangch428

[8] López Marcos, P.; Sanz-Valero, J. (2012). "Presencia y adecuación de los principios activos farmacológicos en la edición española de la Wikipedia". Atención Primaria. **DOI**. Closed access

[9] Vademécum (http://www.webcitation.org/6DJA3sniI). UBM Medica Spain S.A.. Archived from the original (http://www.vademecum.es) on 30 December 2012. Retrieved on 30 December 2012.

[10] DeDeo, S. (2012). Evidence for Non-Finite-State Computation in a Human Social System. *ArXiV*. **PDF** (http://arxiv.org/abs/1212.0018) Open access

[11] http://santafe.edu/~simon

[12] http://sonet.fbk.eu/en/social_networking_group_sonet

[13] Ferron, M. (2012, December 7). *Collective Memories in Wikipedia*. PhD Thesis, University of Trento. **PDF** (http://eprints-phd.biblio.unitn.it/830) Open access

[14] http://www.slideshare.net/fbk.eu/psychological-processes-underlying-wikipedia-representations-of-natural-and-manmade-disasters

[15] Oz, A. (2012). Legitimacy and efficacy: The blackout of Wikipedia. *First Monday*, 17(12). **HTML** (http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/4043/3380) Open access

[16] Normand, J. M.; Sanchez-Vives, M. V.; Waechter, C.; Giannopoulos, E.; Grosswindhager, B.; Spanlang, B.; Guger, C.; Klinker, G. et al. (2012). De Polavieja, Gonzalo G. ed. "Beaming into the Rat World: Enabling Real-Time Interaction between Rat and Human Each at Their Own Scale". PLoS ONE 7 (10): e48331. **DOI**. PMC 3485138. PMID 23118987. Open access

[17] Randall M. Livingstone: Network of Knowledge: Wikipedia as a Sociotechnical System of Intelligence. PDF (https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/12517/Livingstone_oregon_0171A_10498.pdf) Open access

[18] Medeiros, J. (2012). Infographic: History's most influential people, ranked by Wikipedia reach. *Wired UK*. **HTML** (http://www.wired.co.uk/magazine/archive/2012/11/start/wikipedias-top-20-religion-pips-science) Open access

[19] Purcell, K., Rainie, L., Heaps, A., Buchanan, J., Friedrich, L., Jacklin, A., Chen, C., Zickuhr, K. (2012): How Teens Do Research in the Digital World. *Pew Internet* **HTML** (http://www.pewinternet.org/Reports/2012/Student-Research/Summary-of-Findings.aspx) Open access

[20] http://www.pewinternet.org/Reports/2012/Student-Research/Main-Report/Part-3.aspx

[21] Ermann, L., Frahm, K. M., & Shepelyansky, D. L. (2012). Spectral properties of Google matrix of Wikipedia and other networks. *ArXiv* **PDF** (http://arxiv.org/pdf/1212.1068.pdf) Open access

[22] Dobusch, L., & Müller-Seitz, G. (2012). Serial Singularities: Developing a Network Organization by Organizing Events. *Schmalenbach Business Review*, 64, 204–229. **HTML** (http://papers.ssrn.com/abstract=2155083) Open access

[23] Yun, Q., & Gloor, P. A. (2012). The Web Mirrors Value in the Real World – Comparing a Firm's Valuation with Its Web Network Position. *SSRN Electronic Journal*. **DOI** Open access

[24] Morgan, J. T., Bouterse, S., Stierch, S., & Walls, H. (2013). Tea & Sympathy: Crafting Positive New User Experiences on Wikipedia. *CSCW '13*. **PDF** (http://jtmorgan.net/files/morgan_cscw2013_final.pdf) Open access

[25] http://cscw.acm.org/program_papers.html#19

[26] Florin, F., Taraborelli, D., Keyes, O. (2012). Article Feedback: New research and next steps. *Wikimedia blog* **HTML** (http://blog.wikimedia.org/2012/12/20/article-feedback-new-research-and-next-steps/) Open access

[27] Morell, M. F. (2013). Good Faith Collaboration: The Culture of Wikipedia. *Information, Communication & Society*, 16(1), 146–147. **DOI** Closed access

[28] http://reagle.org/joseph/2010/gfc/

[29] http://www.nhm.ac.uk/research-curation/about-science/staff-directory/life-sciences/e-baker/index.html

[30] Baker, E. (2012). Measuring the Impact of Wikipedia for organisations (Part 1), *Ed's blog*, **HTML** (http://pblog.ebaker.me.uk/2012/12/measuring-impact-of-wikipedia-for.html) Open access

Article Sources and Contributors

About *Source*: https://meta.wikimedia.org/w/index.php?oldid=4991282 *Contributors*: DarTar, MZMcBride, Peteforsyth, Piotrus, Rock drum, Tbayer (WMF), Trijnstel, 3 anonymous edits

Issue 2(1): January 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170634 *Contributors*: Daniel Mietchen, DarTar, Gobonobo, Graham87, John of Reading, SMasters, Skomorokh, Tbayer (WMF), Tony1

Issue 2(2): February 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170641 *Contributors*: DarTar, Graham87, Hfordsa, Jodi.a.schneider, Madcoverboy, Skomorokh, Tbayer (WMF), Tony1

Issue 2(3): March 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170646 *Contributors*: Circeus, Daniel Mietchen, DarTar, Graham87, Jodi.a.schneider, Lambiam, Njullien, Piotrus, Skomorokh, Tbayer (WMF), Tony1, Wavelength

Issue 2(4): April 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=5232576 *Contributors*: Amire80, Circeus, Cybercobra, Daniel Mietchen, DarTar, Deor, Funandtrvl, Graham87, Jan Pedersen, Jodi.a.schneider, Lambiam, Nemo bis, Protonk, Steven (WMF), Tbayer (WMF), The ed17, Tony1, 1 anonymous edits

Issue 2(5): May 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170654 *Contributors*: Allens, Charles Matthews, DarTar, Invertzoo, Jodi.a.schneider, Naddy, Piotrus, Scientific29, Tbayer (WMF), The ed17, Tony1, 1 anonymous edits

Issue 2(6): June 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170659 *Contributors*: Adler.fa, Daniel Mietchen, DarTar, Evan (WMF), Graham87, Lambiam, Sgeureka, Siebrand, Tbayer (WMF), The ed17, Tony1, Tpbradbury, 2 anonymous edits

Issue 2(7): July 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170664 *Contributors*: Adler.fa, Benjamin Mako Hill, ChaTo, Circeus, Daniel Mietchen, DarTar, Deor, Gaurav, Jodi.a.schneider, Junkie.dolphin, Mike Peel, Sonia, Tbayer (WMF), The ed17, The wub, Tony1

Issue 2(8): August 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4170667 *Contributors*: AnonMoos, DarTar, Evan (WMF), Graham87, John of Reading, Ragesoss, Shyamal, Tbayer (WMF), The ed17, Tony1, Wdchk

Issue 2(9): September 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=5141058 *Contributors*: AnonMoos, Benjamin Mako Hill, DarTar, EpochFail, GLTester, Jean-Frédéric, Phoebe, Ragesoss, Shyamal, Tbayer (WMF), The ed17, Theopolisme, Tony1

Issue 2(10): October 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4610829 *Contributors*: Adler.fa, AnonMoos, Bdamokos, DarTar, Graham87, Jodi.a.schneider, John of Reading, Orangemike, Phoebe, Ragesoss, Tbayer (WMF), The ed17, Tony1, Wavelength

Issue 2(11): November 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4703865 *Contributors*: Adler.fa, AnonMoos, Benjamin Mako Hill, Circeus, DarTar, John of Reading, Piotrus, Tbayer (WMF), The ed17, Tony1

Issue 2(12): December 2012 *Source*: https://meta.wikimedia.org/w/index.php?oldid=4971037 *Contributors*: Aaronshaw, Adler.fa, Andrew Gray, Benjamin Mako Hill, Cantons-de-l'Est, Daniel Mietchen, DarTar, FallingGravity, Gobonobo, Graham87, Ironholds, John of Reading, Junkie.dolphin, Kaldari, Mervyn, Piotrus, Ragesoss, SoledadKabocha, Tbayer (WMF), The ed17, Tony1, Tpbradbury, 1 anonymous edits

Image Sources, Licenses and Contributors

File:Wikimedia Research Newsletter.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wikimedia_Research_Newsletter.jpg *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: DarTar, Rock drum

File:WRN 2011.pdf *Source*: https://meta.wikimedia.org/w/index.php?title=File:WRN_2011.pdf *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: DarTar, Tbayer (WMF)

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

File:Identica Icon.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Identica_Icon.png *License*: Public Domain *Contributors*: Original uploader was ShakataGaNai at en.wikinews

File:Twitter logo initial.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Twitter_logo_initial.svg *License*: Public Domain *Contributors*: Original uploader was en:User:GageSkidmore, modified by User:Cpro

File:Open Access logo PLoS transparent.svg *Source*: https://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*: https://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:Csueb view.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Csueb_view.jpg *License*: Creative Commons Attribution-Sharealike 2.0 *Contributors*: Jennifer Williams

File:Mind the gap1.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Mind_the_gap1.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: AnonMoos, El Caro, Missvain

File:Wikimedia Foundation Director Sue Gardner speaking at Simmons College on the Wikipedia Public Policy Initiative.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wikimedia_Foundation_Director_Sue_Gardner_speaking_at_Simmons_College_on_the_Wikipedia_Public_Policy_Initiative.jpg *License*: Creative Commons Attribution 3.0 *Contributors*: User:Peteforsyth

File:Keegan, Gergle, & Contractor 2012 Figure 1.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Keegan,_Gergle,_&_Contractor_2012_Figure_1.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:Madcoverboy

File:Pied Piper with Children.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Pied_Piper_with_Children.jpg *License*: Public Domain *Contributors*: Hsarrazin, Jtneill, Kilom691, PhantomS, Túrelio, Wutsje, 4 anonymous edits

File:FlaggedRevs-2-1.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:FlaggedRevs-2-1.svg *License*: Public Domain *Contributors*: Frank Linnenschmidt, derivative SVG Version: hk kng

File:FlaggedRevs-1-1.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:FlaggedRevs-1-1.svg *License*: Public Domain *Contributors*: Frank Linnenschmidt, derivative SVG Version: hk kng

File:Effect of barnstars on productivity.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Effect_of_barnstars_on_productivity.png *License*: Creative Commons Attribution 2.5 *Contributors*: Daniel Mietchen, Timeshifter

File:Wiktionary-logo.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wiktionary-logo.svg *License*: logo *Contributors*: Az1568, Dereckson, FSII, INeverCry, Sevela.p

File:Open Access logo PLoS white.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Open_Access_logo_PLoS_white.svg *License*: Creative Commons Zero *Contributors*: art designer at PLoS, modified by Wikipedia users Nina, Beao, and JakobVoss

File:Harvard Law Review Volume 1.djvu *Source*: https://meta.wikimedia.org/w/index.php?title=File:Harvard_Law_Review_Volume_1.djvu *License*: Public Domain *Contributors*: collective work

File:Goalball vid Paralympics i Aten.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Goalball_vid_Paralympics_i_Aten.jpg *License*: Copyrighted free use *Contributors*: Helene Stjernlöf

File:Newbie quality.by semester.rows.good faith.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Newbie_quality.by_semester.rows.good_faith.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:EpochFail

File:Boxplot of Average Article Feedback ratings by project rated quality.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Boxplot_of_Average_Article_Feedback_ratings_by_project_rated_quality.svg *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:Protonk

File:Intersection13 boundedV2.pdf *Source*: https://meta.wikimedia.org/w/index.php?title=File:Intersection13_boundedV2.pdf *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: Kritik

File:Wikiarthistory.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wikiarthistory.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: Tbayer (WMF)

File:Time evolution of the controversy measure of Michael Jackson.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Time_evolution_of_the_controversy_measure_of_Michael_Jackson.png *License*: Creative Commons Attribution 2.5 *Contributors*: Daniel Mietchen

File:Wiki Loves Monuments 2011 uploads by country.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wiki_Loves_Monuments_2011_uploads_by_country.png *License*: Creative Commons Attribution 3.0 *Contributors*: Daniel Mietchen, ŠJů

File:B&V-B.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:B&V-B.jpg *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: Bill Harry

File:Juvenile bonobo.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Juvenile_bonobo.png *License*: Creative Commons Attribution 2.5 *Contributors*: Hofreiter M, Kreuz E, Eriksson J, Schubert G, Hohmann G

File:Extraction of location, date and taxon data from Field Notes of Junius Henderson on Wikisource.jpeg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Extraction_of_location,_date_and_taxon_data_from_Field_Notes_of_Junius_Henderson_on_Wikisource.jpeg *License*: unknown *Contributors*: User:Gaurav

File:Wikisource 2012 - Aubrey.pdf *Source*: https://meta.wikimedia.org/w/index.php?title=File:Wikisource_2012_-_Aubrey.pdf *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:Aubrey

File:Psychedelic-music-on-wikipedia2.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Psychedelic-music-on-wikipedia2.png *License*: Creative Commons Attribution 3.0 *Contributors*: DarTar, Simon Villeneuve, Tbayer (WMF)

File:Iker Casillas Euro 2012 final trophy.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Iker_Casillas_Euro_2012_final_trophy.jpg *License*: unknown *Contributors*: Bennylin, Simo82, 1 anonymous edits

File:June32009candlevigilHK pic7.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:June32009candlevigilHK_pic7.jpg *License*: Creative Commons Attribution 2.0 *Contributors*: ryanne lai

File:Goddess of Democracy at UBC.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Goddess_of_Democracy_at_UBC.jpg *License*: Public Domain *Contributors*: Original uploader was Dr.kwan at en.wikipedia

File:Desirable newcomer quality over time.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Desirable_newcomer_quality_over_time.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:EpochFail

File:Desirable newcomer reverts over time.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Desirable_newcomer_reverts_over_time.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:EpochFail

File:Desirable newcomer survival over time.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Desirable_newcomer_survival_over_time.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:EpochFail

File:Rorschach blot 10.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Rorschach_blot_10.jpg *License*: Public Domain *Contributors*: Hermann Rorschach (died 1922)

File:Democrats_and_Republicans_in_Wikipedia_discussions_green.png *Source*: https://meta.wikimedia.org/w/index.php?title=File:Democrats_and_Republicans_in_Wikipedia_discussions_green.png *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:Aldivad

Image:Graph betweenness.svg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Graph_betweenness.svg *License*: Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors*: Claudio Rocchini

File:Medication potofen(Ibuprofen).JPG *Source*: https://meta.wikimedia.org/w/index.php?title=File:Medication_potofen(Ibuprofen).JPG *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: Mk2010

File:Flickr - Floris Van Cauwelaert - The messages on Tahrir Square (8).jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Flickr_-_Floris_Van_Cauwelaert_-_The_messages_on_Tahrir_Square_(8).jpg *License*: Creative Commons Attribution-Sharealike 2.0 *Contributors*: Floris Van Cauwelaert from Brussels, Belgium

File:History Wikipedia English SOPA 2012 Blackout2.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:History_Wikipedia_English_SOPA_2012_Blackout2.jpg *License*: unknown *Contributors*: Pseudoanonymous at en.wikipedia

File:Beaming-into-the-Rat-World-Enabling-Real-Time-Interaction-between-Rat-and-Human-Each-at-Their-Own-pone.0048331.s006.ogv *Source*: https://meta.wikimedia.org/w/index.php?title=File:Beaming-into-the-Rat-World-Enabling-Real-Time-Interaction-between-Rat-and-Human-Each-at-Their-Own-pone.0048331.s006.ogv *License*: Creative Commons Attribution 2.5 *Contributors*: Normand J, Sanchez-Vives M, Waechter C, Giannopoulos E, Grosswindhager B, Spanlang B, Guger C, Klinker G, Srinivasan M, Slater M

File:Googlematrixwikipedia2009.jpg *Source*: https://meta.wikimedia.org/w/index.php?title=File:Googlematrixwikipedia2009.jpg *License*: GNU Free Documentation License *Contributors*: Bulwersator

File:AFT5 2012-Q4 report.pdf *Source*: https://meta.wikimedia.org/w/index.php?title=File:AFT5_2012-Q4_report.pdf *License*: Creative Commons Attribution-Sharealike 3.0 *Contributors*: User:DarTar

License