

Wikimedia Research Newsletter

Volume 4 (2014)

Contents

1 About	1
1.1 Facts and figures	1
1.2 How to subscribe	1
1.3 How to contribute	2
1.4 Open access vs. closed access publications	2
1.5 Archives	3
1.5.1 Volume 6 (2016)	3
1.5.2 Volume 5 (2015)	3
1.5.3 Volume 4 (2014)	3
1.5.4 Volume 3 (2013)	3
1.5.5 Volume 2 (2012)	3
1.5.6 Volume 1 (2011)	4
1.5.7 Search the WRN archives	4
1.6 Contact	4
2 Issue 4(1): January 2014	5
2.0.1 Translation students embrace Wikipedia assignments, but find user interface frustrating	5
2.1 Briefly	6
2.1.1 References	7
3 Issue 4(2): February 2014	8
3.0.2 CSCW '14 retrospective	8
3.0.3 Clustering Wikipedia editors by their biases	9
3.0.4 Monthly research showcase launched	9
3.0.5 Study of AfD debates: Did the SOPA protests mellow deletionists?	9
3.0.6 Word frequency analysis identifies “four conceptualisations of femininity on Wikipedia”	10
3.0.7 Wikipedia and the development of academic language	11
3.0.8 Briefly	11
3.0.9 References	12
4 Issue 4(3): March 2014	14
4.0.10 Cross-language study of conflict on Wikipedia	14
4.0.11 The social construction of knowledge on English Wikipedia	14

4.0.12	User hierarchy map: Building Wikipedia's Org Chart	15
4.0.13	Briefly	15
4.0.14	References	17
5	Issue 4(4): April 2014	18
5.0.15	Wikipedia Usage Estimates Prevalence of Influenza-Like Illness	18
5.0.16	Survey of academics' view on Wikipedia and open-access publishing	18
5.0.17	Briefly	19
5.0.18	Other recent publications	19
5.0.19	References	20
6	Issue 4(5): May 2014	22
6.0.20	"Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership"	22
6.0.21	Chinese-language time-zones favor Asian pop and IT topics on Wikipedia	22
6.0.22	"Bipartite editing prediction in Wikipedia"	23
6.0.23	Briefly	23
6.0.24	Other recent publications	24
6.0.25	References	25
7	Issue 4(6): June 2014	26
7.0.26	New book: <i>Global Wikipedia</i>	26
7.0.27	"Interactions of cultures and top people of Wikipedia from ranking of 24 language editions"	26
7.0.28	"Recommending reference materials in context to facilitate editing Wikipedia"	27
7.0.29	"What do Chinese-language microblog users do with Baidu Baike and Chinese Wikipedia?"	28
7.0.30	Content or people? Achieving critical mass to promote growth in WikiProjects	28
7.0.31	"Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity"	28
7.0.32	Briefly	29
7.0.33	Other recent publications	29
7.0.34	References	29
8	Issue 4(7): July 2014	31
8.0.35	Understanding shifting values underlying the paid content debate on the English Wikipedia	31
8.0.36	"Pivot-based multilingual dictionary building using Wiktionary"	31
8.0.37	"Cross Language Learning from Bots and Users to detect Vandalism on Wikipedia"	32
8.0.38	Readers' interests differ from editors' preferences	33
8.0.39	Wikipedia from the perspective of PR and marketing	33
8.0.40	"No praise without effort: experimental evidence on how rewards affect Wikipedia's contributor community"	34
8.0.41	Briefly	34
8.0.42	Other recent publications	34
8.0.43	References	35
9	Issue 4(8): August 2014	37

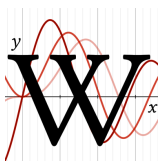
9.0.44	Wikipedia in all languages used to rank global historical figures of all time	37
9.0.45	WikiBrain: Democratizing computation on Wikipedia	38
9.0.46	Newcomer productivity and pre-publication review	39
9.0.47	Briefly	40
9.0.48	References	41
10	Issue 4(9): September 2014	42
10.0.49	“Reliability of user-generated data: the case of biographical data in Wikipedia”	42
10.0.50	Focused Wikipedians stay active longer	42
10.0.51	“WordNet-Wikipedia-Wiktionary: construction of a three-way alignment”	43
10.0.52	Briefly	43
10.0.53	Other recent publications	44
10.0.54	References	45
11	Issue 4(10): October 2014	46
11.0.55	Tl;dr: Users, informed consent and privacy policies online	46
11.0.56	Briefly	46
11.0.57	Other recent publications	48
11.0.58	References	48
12	Issue 4(11): November 2014	50
12.0.59	“Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia”	50
12.0.60	In nutritional articles, academic citations rise while news media citations decrease	51
12.0.61	Wikipedia user session timing compared with other online activities	51
12.0.62	Briefly	52
12.0.63	Other recent publications	53
12.0.64	References	54
13	Issue 4(12): December 2014	55
13.0.65	Use of Wikipedia in higher education influenced by peer opinions and perception of Wikipedia’s quality	55
13.0.66	Analysis of two gender-driven talk page conflicts on the German-language Wikipedia	55
13.0.67	Briefly	56
13.0.68	Other recent publications	57
13.0.69	References	57
13.1	Text and image sources, contributors, and licenses	59
13.1.1	Text	59
13.1.2	Images	59
13.1.3	Content license	60

Chapter 1

About

- [home](#)
- [projects](#)
- [support](#)
- [data](#)
- [resources](#)
- [newsletter](#)
- [committee](#)
- [chat](#)

Wikimedia Research Newsletter



[Home](#) • [Latest issue: January 2016](#)[\[contribute\]](#) [\[archives\]](#)



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**.

1.1 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**, 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** or downloaded (as a BibTeX, RIS, PDF file or in other formats), ready to be imported into reference managers or other bodies of wiki research literature.[Read more...](#)

The twelve issues of the **second volume** (January-December 2012) covered 225 publications. This corpus can be **browsed online** on Zotero, or downloaded as BibTeX file from datahub.io. [Read more...](#)

1.2 How to subscribe

- To receive the full text of each new issue in the form of an **HTML email**, [sign up here](#).
- You can also subscribe to the newsletter on the **Wikimedia Foundation's blog** via the following RSS feed: 
- The table of contents of each issue is cross-posted to the **wiki-research-l** mailing list.

RECENT RESEARCH

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

By [Junkie.dolphin](#), [Lilarioa](#), [Haeb](#), [DarTar](#), [Steven Walling](#), [Daniel Mischen](#), 25 July 2011 [Share this](#) [Edit](#)

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

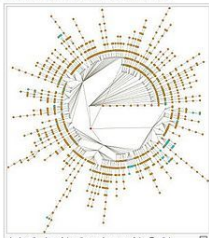
A study covered in the [previous edition](#) of the research newsletter was extended and published by the authors on [ArXiv](#). The authors report a new method for classifying how disputed a Wikipedia article is, to detect controversies and edit-wars. At its core, the method is based on looking at pairs 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 [Istanbul](#) 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 'assassination of Barack Obama', displaying the article as the root node in red, "talk" nodes in yellow, structural elements of the discussion page in gray, and unsigned comments in blue. From Laniado et al., 2011. ^[3]

The paper moves on to examine the degree [associativity](#) 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 [disassortivity](#), 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 contentfulness 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.

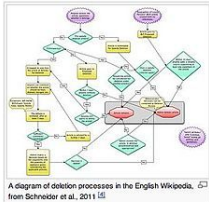
Two of the authors gave a presentation at last month's [Hypertext 2011](#) conference in Eindhoven: "[Co-authorship 2.0: patterns of collaboration in Wikipedia](#)".

- [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. ^[4] 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 a user's [win/loss](#) (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 [BDE](#) 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 [AID](#) discussions and illustrates with a diagram the complexity of deletion discussions and procedures in the English Wikipedia. ^[6]



A diagram of deletion processes in the English Wikipedia, from Schreiber et al., 2011. ^[6]

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

- Follow the [@WikiResearch](#) feed on [Twitter](#). 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.

1.3 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](#), 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 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](#) (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.

1.4 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**.

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** .

1.5 Archives

1.5.1 Volume 6 (2016)

- WRN 6(01) – January 2016

1.5.2 Volume 5 (2015)

- WRN 5(12) – December 2015
- WRN 5(11) – November 2015
- WRN 5(10) – October 2015
- WRN 5(9) – September 2015
- WRN 5(8) – August 2015
- WRN 5(7) – July 2015
- WRN 5(6) – June 2015
- WRN 5(5) – May 2015
- WRN 5(4) – April 2015
- WRN 5(3) – March 2015
- WRN 5(2) – February 2015
- WRN 5(1) – January 2015

1.5.3 Volume 4 (2014)

- WRN 4(12) – December 2014
- WRN 4(11) – November 2014
- WRN 4(10) – October 2014
- WRN 4(9) – September 2014
- WRN 4(8) – August 2014
- WRN 4(7) – July 2014
- WRN 4(6) – June 2014
- WRN 4(5) – May 2014
- WRN 4(4) – April 2014
- WRN 4(3) – March 2014
- WRN 4(2) – February 2014
- WRN 4(1) – January 2014

1.5.4 Volume 3 (2013)

- WRN 3(12) – December 2013
- WRN 3(11) – November 2013
- WRN 3(10) – October 2013
- WRN 3(9) – September 2013
- WRN 3(8) – August 2013
- WRN 3(7) – July 2013
- WRN 3(6) – June 2013
- WRN 3(5) – May 2013
- WRN 3(4) – April 2013
- WRN 3(3) – March 2013
- WRN 3(2) – February 2013
- WRN 3(1) – January 2013

1.5.5 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

1.5.6 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

1.5.7 Search the WRN archives

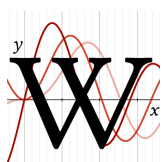
1.6 Contact

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

Chapter 2

Issue 4(1): January 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 1 • January 2014 [\[contribute\]](#) [\[archives\]](#)



Translation assignments, weasel words, and Wikipedia's content in its later years

With contributions by: Aaron Halfaker, Jonathan Morgan, Piotr Konieczny and Tilman Bayer

2.0.1 Translation students embrace Wikipedia assignments, but find user interface frustrating

An article, “Translating Wikipedia Articles: A Preliminary Report on Authentic Translation Projects in Formal Translator Training”, ^[1] reports on the author's experiment with “a promising type of assignment in formal translator training which involves translating and publishing Wikipedia articles”, in three courses with second- and third-year students at the Institute of English Studies, University of Warsaw.

It was “enthusiastically embraced by the trainees ... Practically all of the respondents [in a participant survey] concluded that the experience was either 'positive' (31 people, 56% of the respondents) or 'very positive' (23 people, 42% of the respondents).” And “more than 90% of the respondents (50 people) recommended that the exercise 'should definitely be kept [in future courses], maybe with some improvements,' and the remaining 5 people (9%) cautioned that improvements to the format were needed before it was used again. No-one recommended culling the exercise from the syllabus.”

However, the author cautions that Polish–English translations required more instructor feedback and editing than translations from English into Polish (the students' native language). And “most people found the technological aspects of the assignment frustrating, with most students assessing them as either 'hard' (39%) or 'very hard' (16%) to complete. The technical skills involved not only coding and formatting using Wikipedia's idiosyncratic syntax, but the practical aspects of publication. [Asked] to identify areas requiring better assistance, the respondents predominantly focused on the need for better information on coding/formatting the article and on publishing the entry. Thirty-nine people (almost three-quarters of the respondents) found the publication criteria baffling enough to postulate that more assistance was needed. That is even more than the 36 people (68%) who had problems dealing with Wikipedia's admittedly idiosyncratic code.”

In the researcher's observation, this contributed to the initially disappointing success rate: “Of the 59 respondents, only eight had their work accepted [after drafting it in a sandbox]. Seven people were asked to revise their entries to bring them into line with Wikipedia's publication guidelines but neglected to do so, and 36 did not even try to publish. Some of those people were still waiting for their feedback to get a green light, but this result can only be described as a big disappointment. ... After a resource pack on how to translate and publish a Wikipedia entry was distributed to a fresh batch of students in the following semester, the successful publication rate proved significantly higher.” These English-language instructions are humorously written in the form of a game manual (“Your mission is to create a Polish translation of an English-language article and deliver it safely to the Free Encyclopaedia HQ officially known as 'Wikipedia'. Sounds easy? Think again. Wikipedia is defended by an army of Editors who guard its gates night and day to stop Lord Factoid and his minions from corrupting it with bad articles.”). They are available on the author's website, together with a small list of the resulting articles (which is absent from the actual research paper).

The project was inspired by author Cory Doctorow's use of Wikipedia in a 2009 course – most likely the one listed [here](#), although the paper fails to specify it. The absence of discussion of the Wikipedia policies, combined with

the absence of any references to prior research from the field of Wikipedia in education, makes it almost certain that the author was unaware of Wikipedia policies and available support (Wikipedia Education Program, etc.).

2.1 Briefly

Why bots should be regarded as an integral part of Wikipedia's software platform

In a new paper titled "Bots, bespoke code, and the materiality of software platforms"^[2] published in *Information, Communication & Society*, Stuart Geiger (User:Staeiou) presents a critical reflection on the common view of online communities as sovereign platforms governed by code, using Wikipedia as an example. He borrows the term "bespoke" to refer to code that affects the social dynamics of a community, but is designed and owned separately from the software platform (e.g. Wikipedia bots). Geiger mixes vignettes describing his personal experience running with discussions of the related literature (including Lessig's famous "code is law") to advocate "examining online communities as both governed by stock and bespoke code, or else we will miss important characteristics of mediated interaction."

"Precise and efficient attribution of authorship of revised content"

Using a graph-theoretic approach, Flöck and Acosta investigate^[3] a new algorithm that can detect the author of a part of document that has been edited by many. They use a units-of-discourse model, to identify paragraphs, sentences and words, and their connections. The authors claim that this approach can identify an author with 95% precision, which is more than the current state-of-the-art. Most intriguing is that to make this comparison they have created the first "gold standard", a hand-made benchmark of 240 Wikipedia pages and their complex authorship histories.

"Which news organizations influence Wikipedia?"

This is the question asked in a blog post^[4] by a post-doc researcher at Columbia University's Tow Center for digital journalism. Looking at the top 10 news stories of 2013 – an admittedly subjective set determined by the author – the organizations from which the citations come are analyzed. Leading the pack are the *New York Times*, *Washington Post* and CNN, but the author notes that the tail of the distribution is very long – 68% of citations are not produced by the top 10 organizations. Qualitative analysis discusses "the surprise for the news organizations that don't make the top ten; CBS News, ABC News, FOX News [...] this top ten strikes as leaning left overall".

Weasels, hedges, and peacocks in Wikipedia articles

Some computational linguists find many Wikipedia articles to be a superlative corpus for natural language processing applications. Weasel words, hedges, and peacock terms (like the ones in the previous sentence) are labelled by Wikipedia editors because they tend to make an article less objective. A recent study^[5] leverages this work to understand general features of the way people use subjective language to increase uncertainty about the truth or authority of the statements they make. By examining a set of 200 Wikipedia articles that had been flagged for these terms, the researchers found 899 different keywords that were frequently used as peacock terms, weasel words, and hedges. A machine learning classifier that was trained on this set of key words was able to identify other (unlabelled) articles that were written in a subjective manner, with high accuracy. In the future, approaches like these could lead to better automated detection of inappropriately subjective or unsourced statements—not only in Wikipedia articles, but also news articles, scientific papers, product reviews, search results, and other scenarios where people need to be able to trust that the information they are reading is credible.

WikiSym/OpenSym call for submissions

The call for submissions (until April 20) to this year's WikiSym/OpenSym conference lists 15 research topics of interest in the Wikipedia research track. The conference has taken place annually since 2005; this year's instance will take place from August 27–29, 2014 in Berlin, Germany. As in preceding years, the organizers intend to apply for financial support from the Wikimedia Foundation, addressing the open access concerns voiced in previous years with a reference to a new policy of ACM, the publisher of the proceedings.

Gender imbalance in Wikipedia coverage of academics to be studied with 2-year NSF grant

Sociologists Hannah Brückner (New York University Abu Dhabi) and Julia Adams (Yale University) have received a two-year grant over US\$132,000 from the National Science Foundation for a research project titled "Collaborative Research: Wikipedia and the Democratization of Academic Knowledge". As described in a press release this month, the project will study "the way gender bias affects the development of pages for American academics in the fields of computer science, history, and sociology, disciplines that vary in their gender composition. ... For instance, 80 percent of academics listed on the Wikipedia page American Sociologists are male, while in reality less than 60 percent of American sociologists are male." The researchers plan to create lists of academics in each field who satisfy the notability criteria for academics, and compare them with the actual cover-

age on Wikipedia.

Discussions about accessibility studied

A paper presented at last year's *SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*^[6] examines the English Wikipedia as one of two “case studies of two UGC communities with accessible content”. Starting from uses of `Template:AccessibilityDispute`, and pages related to `Wikipedia:WikiProject_Accessibility`, the authors “identified 179 accessibility discussions involving 82 contributors” and coded them according to content and other aspects.

Wikipedia content “still growing substantially even in later years”

A preprint^[7] by two researchers from Stanford University and the London School of Economics analyzes the history of around 1500 pages in the English Wikipedia's `Category:Roman Empire` over eight years, providing descriptive statistics for 77,671 (non-bot) edits for articles in that category. The authors find that “content is still growing substantially even in later years. Less new pages are created over time, but at the page-level we see very little slow-down in activity.” They identify a “key driver of content growth which is a spill-over effect of past edits on current editing activity” – that is, articles that have been edited more often in the past attract more editing activity in the future, even when controlling for factors such as the page's “inherent popularity”, suggesting a causal relationship.

Discover “winning arguments” in article histories, and notify losing editors

Winning the best paper award at last year's European Semantic Web Conference (ESWC), three authors from the French research institute INRIA presented (video)^[8] “a framework to support community managers in managing argumentative discussions on wiki-like platforms. In particular, our approach proposes to automatically detect the natural language arguments and the relations among them, i.e., support or challenges, and then to organize the detected arguments in bipolar argumentation frameworks.” Specifically, they analyzed the revision history of the five most revised pages on the English Wikipedia at one point (e.g. George W. Bush), extracting sentences that were heavily edited over time while still describing the same event. To these “arguments” they apply a NLP technique known as textual entailment (basically, detecting whether the assertion of the new version of the sentence logically follows from the first version, or whether the first version was “attacked” by a subsequent editor by deleting or correcting some of the information). The paper focuses mostly on establishing and testing this

methodology, without detailing the actual results derived from the five revision histories (i.e. which arguments actually won in those cases), but the authors promise that “this kind of representation helps community managers to understand the overall structure of the discussions and which are the winning arguments.” Also, they point out that it should make it possible to “notify the users when their own arguments are attacked.”



2.1.1 References

- [1] Piotr Szymczak: Translating Wikipedia Articles: A Preliminary Report on Authentic Translation Projects in Formal Translator Training. In: *Acta Philologica* 44 (Warszawa 2013) <http://acta.neofilologia.uw.edu.pl/archiwum/acta44.pdf> p.61ff
- [2] Geiger, R. Stuart. “Bots, bespoke code, and the materiality of software platforms”. *Information, Communication & Society*: 1–15. doi:10.1080/1369118X.2013.873069. ISSN 1369-118X. , author's copy at <http://stuartgeiger.com/bespoke-code-ics.pdf>
- [3] Fabian Flöck, Maribel Acosta: WikiWho: Precise and Efficient Attribution of Authorship of Revisioned Content. <http://www.aifb.kit.edu/web/Inproceedings3398>
- [4] Fergus Pitt: Which News Organizations Influence Wikipedia? January 17, 2014, <http://towcenter.org/blog/which-news-organizations-influence-wikipedia/>
- [5] Vincze, Veronika: Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. <http://www.aclweb.org/anthology/I113/I13-1044.pdf>
- [6] Kuksenok, Katie; Brooks, Michael; Mankoff, Jennifer (2013). “Accessible Online Content Creation by End Users”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York City: ACM. pp. 59–68. doi:10.1145/2470654.2470664. ISBN 978-1-4503-1899-0.
- [7] Aleksu Aaltonen, Stephan Seiler: Cumulative Knowledge and Open Source Content Growth: The Case of Wikipedia http://faculty-gsb.stanford.edu/seiler/documents/wiki_dec2013_03.pdf
- [8] Elena Cabrio, Serena Villata, and Fabien Gandon: A Support Framework for Argumentative Discussions Management in the Web. <http://eswc-conferences.org/sites/default/files/papers2013/cabrio.pdf>

Wikimedia Research Newsletter

Vol: 4 • Issue: 1 • January 2014

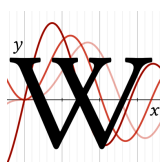
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost

Subscribe:  [Email](#)  • [archives] [signpost edition] [contribute] [research index]

Chapter 3

Issue 4(2): February 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 2 • February 2014 [contribute] [archives]



CSCW '14 retrospective; the impact of SOPA on deletionism; like-minded editors clustered; Wikipedia stylistic norms as a model for academic writing

With contributions by: David Ludwig, Morten Warncke-Wang, Maximilian Klein, Piotr Konieczny, Giovanni Luca Ciampaglia, Dario Taraborelli and Tilman Bayer

3.0.2 CSCW '14 retrospective

The *17th ACM Conference on Computer-supported cooperative work and Social Computing* (CSCW '14) took place this month in Baltimore, Maryland.^[supp 1] The conference brought together more than 500 researchers and practitioners from industry and academia presenting research on “the design and use of technologies that affect groups, organizations, communities, and networks.” Research on Wikipedia and wiki-based collaboration has been a major focus of CSCW in the past. This year, three papers on Wikipedia were presented:

The rise of alt.projects in Wikipedia

Jonathan Morgan from the Wikimedia Foundation and collaborators from the University of Washington^[1] analyzed the nature of collaboration in *alternative* WikiProjects, i.e. projects that the authors identify as not following “the conventional pattern of coordinating a loosely defined range of article creation and curation-related ac-

tivities *within a well defined topic area*” (examples of such alternative WikiProjects include the Guild of Copy Editors or WikiProject Dispute Resolution). The authors present an analysis of editing activity by members of these projects that are not focused on topic content editing. The paper also reports data on the number of contributors involved in WikiProjects over time: while the number of editors participating in conventional projects decreased by 51% between 2007 and 2012, participation in alternative projects only declined by 13% in the same period and saw an overall 57% increase in the raw number of contributions.

Categorizing barnstars via Mechanical Turk

Paul Andre and collaborators from Carnegie Mellon University presented a study showing how to effectively crowdsource a complex categorization task by assigning it to users with no prior knowledge or domain expertise.^[2] The authors selected a corpus of Wikipedia barnstars and showed how different task designs can produce crowdsourced judgments where Mechanical Turk workers accurately match expert categorization. Expert categorization was obtained by recruiting two Wikipedians with substantial editing activity as independent raters.

Understanding donor behavior through email

A team of researchers from Yahoo! Research, the Qatar Computing Research Institute and UC Berkeley analyzed two months of anonymized email logs to understand the demographics, personal interests and donation behavior of individuals responding to different fundraising campaigns.^[3] The results include donation email from the Wikimedia Foundation and indicate that among other campaigns, email from a *wikimedia.org* domain had the highest score of messages tagged for spam over total messages read, which the authors attribute to spoofing. The paper also indicates that the Wikimedia fundraiser tends to attract slightly more male than female donors.

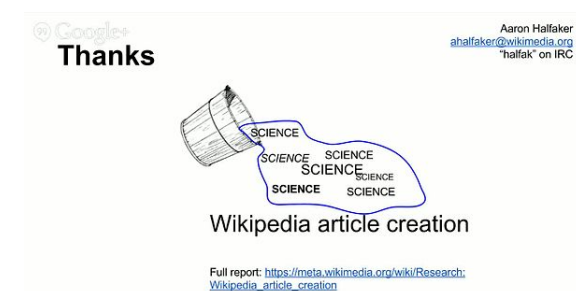
3.0.3 Clustering Wikipedia editors by their biases

review by User:Maximilianklein

Building on the streams of rating editors by *content persistence* and algorithmically finding *cliques* of editors, Nakamura, Suzuki and Ishikawa propose^[4] a sophisticated tweak to find like- and disparate-minded editors, and test it against the Japanese Wikipedia. The method works by finding cliques in a **weighted graph** between all editors of an article and weighting the edges by the agreement or disagreement between editor. To find the agreement between two editors, they iterate through the full edit history and use the **content persistence axioms** of interpreting edits that are leaving text unchanged as agreement, and deleting text as disagreement. Addressing that leaving text unchanged is not always a strong indication of agreement, they normalize by each action's frequency of both the source editor and the target editor. That is, the method accounts for the **propensity** of an editor to change text, and the propensity of editors to have their text changed.

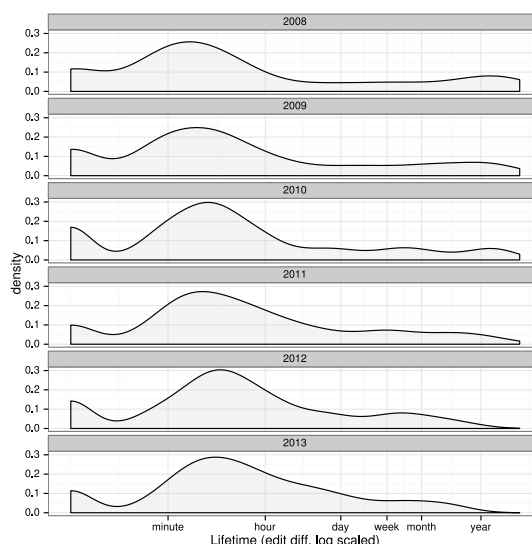
To verify their method, its results are compared to a simplified weighting scheme, random clustering, and human-clustered results on 7 articles in Japanese Wikipedia. In 6 out of 7 articles, the proposed technique beats simplified weighting. An example they present is their detection of pro- and anti-nuclear editors on the **Nuclear Power Plant** article. An implication of such detection would be a gadget that colours text of an article depending on which editor group wrote it.

3.0.4 Monthly research showcase launched



Video of the February 2014 Research Showcase

The Wikimedia Foundation's Research & Data team announced its first public showcase, a monthly review of work conducted by researchers at the Foundation. Aaron Halfaker presented a study of trends in newcomer article creation across 10 languages with a focus on the English and German Wikipedias (slides). The study indicates that in wikis where anonymous users can create articles, their articles are less likely to be deleted than articles created by newly registered editors. Oliver Keyes presented an anal-



The lifetime of deleted articles by year of creation

ysis of how readers access Wikipedia on mobile devices and reviewed methods to identify the typical duration of a **mobile browsing session** (slides). The showcase is hosted at the Wikimedia Foundation every 3rd Wednesday of the month and live streamed on YouTube.^[supp 2]

3.0.5 Study of AfD debates: Did the SOPA protests mellow deletionists?



Wikipedia's SOPA blackout

A paper titled "What influences online deliberation? A wikipedia [sic] study"^[5] studies rationales used by participants in deletion discussions, in the larger context of democratic online deliberation. The authors reviewed in detail deletion discussions for a total of 229 articles, listed for deletion on three dates, one of them being January 15th, 2012, three days prior to the the English Wikipedia's global blackout as part of the Wikipedia: SOPA initiative. The authors looked into whether this event would influence rationales of the deletion discussions and their outcome. They also reviewed, in less detail, a number of other deletions from around the time of the SOPA protest. The authors display a good knowledge of relevant literature, including that in the field of Wikipedia studies, presenting an informative literature

review section.

Overall, the authors find that the overall quality of the discussions is high, as most of the participants display knowledge of Wikipedia's policies, particularly on the notability and credibility (or what we would more likely refer to as reliability) of the articles whose deletion is considered. In re, notability far outweighs the second most frequent rationale, credibility (reliability). They confirm that the deletion system works as intended, with decisions made by majority voters.

Interestingly, the authors find that certain topics did tend to trigger more deletion outcomes, said topics being articles about people, for-profit organizations, and definitions. In turn, they observe that "locations or events are more likely to be kept than expected, and articles about nonprofit organizations and media are more likely to be suggested for other options (e.g., merge, redirect, etc.) than expected". Discussions about people and for-profit organizations were more likely to be unanimous than expected, whereas articles about nonprofit organizations, certain locations, or events were more likely to lead to a non-unanimous discussion. Regarding the SOPA protests' influence on deletion debates, the authors find a small and short-lived increase in keep decisions following the period of community mobilization and discussion about the issue, and tentatively attribute this to editors being impacted by the idea of internet freedom and consequently allowing free(er) Internet publishing.

The authors sum up those observations, noting that "the community members of Wikipedia have clear standards for judging the acceptability of a biography or commercial organization article; and such standards are missing or less clear when it comes to the topics on location, event, or nonprofit organization ... Thus, one suggestion to the Wikipedia community is to make the criteria of judging these topics more clear or specific with examples, so it will alleviate the ambiguity of the situation". This reviewer, as a participant of a not insignificant number of deletion discussions as well as those about the associated policies, agrees with said statement. With regards to the wider scheme, the authors conclude that the AfD process is an example of "a democratic deliberation process interested in maintaining information quality in Wikipedia".

3.0.6 Word frequency analysis identifies "four conceptualisations of femininity on Wikipedia"

In a linguistics student paper^[6] at Lund University, the author reviews the linguistic conceptualisation of femininity on (English) Wikipedia, with regards to whether language used to refer to women differs depending on the type of articles it is used in. Specifically, the author analyzed the use of five lexemes (a term which in the context of this study means words): *ladylike*, *girly*, *girlish*, *feminine* and *womanly*. The findings confirm that the usage of those



Girl with Cherries by Ambrogio de Predis (the current lead illustration of the article femininity)

terms is non-accidental. The word *feminine*, most commonly used of the five studied, correlates primarily to the topics of fashion, sexuality, and to a lesser extent, culture, society and female historical biographies. The second most popular is the word *womanly*, which in turn correlates with topics of female artists, religion and history. *Girlish*, the fourth most popular word, correlates most strongly with the biographies of males, as well as with the articles on movies and TV, female entertainers, literature and music. Finally, *girly* and *ladylike*, respectively 3rd and 5th in terms of popularity, cluster together and correlate to topics such as movies and TV (animated), Japanese culture, art, tobacco and female athletes. Later, the author also suggests that there is a not insignificant overlap in usage between the cluster for *girlish* and the combined cluster for *girly* and *ladylike*. He concludes that there are three or four different conceptualisations of femininity on Wikipedia, which in more simple terms means, to quote the author, that "people do indeed represent women in different ways when talking about different things [on Wikipedia]", with "*girly* and *girlish* having a somewhat frivolous undertone and *womanly*, *feminine* and *ladylike* being of a more serious and reserved nature".

The study does suffer from a few issues: a literature review could be more comprehensive (the paper cites only six works, and not a single one of them from the field of Wikipedia studies), and this reviewer did not find sufficient justification for why the author limited himself to the analysis of only 500 occurrences (total) of the five lexemes studied. A further discussion of how the said 500 cases were selected would likely strengthen the paper.

3.0.7 Wikipedia and the development of academic language

Ursula Reutner's article "Wikipedia und der Wandel der Wissenschaftssprache"^[7] discusses Wikipedia's linguistic norms and style as a case study for the development of academic language.

The article is divided into three main sections. After providing some historical context about Wikipedia and the history of encyclopedias (section 1), the article focuses on linguistic norms in Wikipedia and their relation to linguistic norms in academic language (section 2). Reutner identifies five crucial linguistic norms in Wikipedia: (1) non-personal language such as the avoidance of first- and second-person pronouns, (2) neutral language as expressed in the policy of a "neutral point of view", (3) avoidance of redundancies, (4) avoidance of unnecessarily complex wording, and (5) focus on simple syntax and the use of short independent clauses. Although Reutner mentions many well-known differences between Wikipedia and traditional forms of academic writing (e.g. the dynamic, collaborative, and partly non-academic character of Wikipedia), she stresses that the policies of Wikipedia largely follow traditional norms of academic writing.

The third section focuses on case studies of Wikipedia articles (mostly *fr:Euro* and *it:Euro*) and finds a large variety of norm violations that suggest a gap between linguistic norms and actual style in Wikipedia. Reutner's examples of biased, clumsy, and long-winded formulations hardly come as a surprise as these quality issues are well-known topics in Wikipedia research^[supp 3]. However, Reutner's analysis is not limited to quality problems but also addresses further interesting features of Wikipedia articles. For example, she points out that Wikipedia differs from many print encyclopedias in Romanic languages such as the *Grande Dizionario Enciclopedico* (1964) or the *Enciclopedia Treccani* (2010) through a focus on accessibility as illustrated by the use of copular sentences at the beginning of articles and the repetition of crucial ideas and terms. Furthermore, Reutner argues that Wikipedia differs from other forms of academic writing through narrative elements and a generous use of space.

Reutner's findings raise general questions regarding the relation between Wikipedia and the development of academic language and her short conclusion makes three suggestions: First, Wikipedia's policies largely follow traditional norms of academic writing. Second, the digital, collaborative, and partly non-academic character of Wikipedia leads to "emotional and dialogic elements that are surprising in the tradition of encyclopedias" (p.17). Third, the focus on accessibility follows an Anglo-American tradition of academic writing (even in the Italian and French language versions). Although Reutner's conclusions seem well-justified, they leave the question open whether Wikipedia reflects or even influences the

general development of academic language. For example, one may argue that many of Reutner's findings are effects of the partly *non-academic* character of Wikipedia and therefore not representative of the development of academic language. Other linguistic features are arguably effects of collaborative text production and it would be interesting to compare Reutner's findings with other collaborative and non-collaborative forms of academic writing. Finally, one may worry that some of Reutner's findings are artifacts of a small and biased sample. For example, Reutner only considers articles (*de:Euro*, *en:Euro*, *es:Euro*, *fr:Euro*, and *it:Euro*) that are created by large and diverse author groups but does not discuss more specialized articles that usually only have one or two main authors. However, it is well-known that the style and quality of Wikipedia articles depends on variables such as group size and group composition^[supp 4] and diverse forms of collaboration patterns^[supp 5]. It would therefore be interesting to discuss Reutner's linguistic findings in the context of a more diverse sample of Wikipedia articles.

3.0.8 Briefly

Wikipedia's assessability

A paper to be presented at the upcoming *Conference on Human Factors in Computing Systems* (CHI '14)^[8] by Forte, Andalibi, Park, and Willever-Farr introduces a vocabulary for "assessable design". Their framework considers social and technological approaches to information literacy in combination with consumption and production. From interviewing Wikipedians, librarians, and novices about their understanding of Wikipedia articles, the authors identify two important concepts of assessable design: provenance and stewardship. The authors then test these concepts in an experiment, finding that exposing readers to these can have large effects on their assessment of not only articles but Wikipedia as a whole. Considering whether their framework can be generalized to the assessability of content on other informational websites, the authors caution that "Wikipedia is a remarkably conservative resource given its reputation as a renegade reference. Policies surrounding citation defer to well-established publishing processes like scientific peer review and traditional journalism and prohibit the production of personalized content."

"Finding missing cross-language links in Wikipedia"

This is the title of a paper^[9] in the *Journal of Information and Data Management*. Using a combination of feature extraction and a decision tree classifier, the authors seek to discover missing inter-language links (ILL) between the English and Portuguese Wikipedia editions. The authors hypothesise that there are roughly 165,000 missing ILLs in each of the Wikipedias, but do not appear to take previous research on the overlap of Wikipedia con-

tent into consideration.^[supp 6] Two novel features are introduced: category linking and ILL transitivity. Performance is evaluated using a dataset of known connected and disconnected articles where the French, Italian, and Spanish Wikipedias are used as intermediate languages for discovering link transitivity. Category linking is identified as a useful way of discovering candidate articles for linking, while link transitivity is the key feature for correctly identifying links. Today, Wikidata's central repository of ILLs makes link transitivity mostly a moot problem, but that is not addressed by the authors.

“Spillovers in Networks of User Generated Content”

A discussion paper^[10] by economists at the Centre for European Economic Research (ZEW) reports an analysis of content curation and consumption under spikes of attention. The authors analyzed 23 examples of pages that underwent a sudden surge of attention, either because they were featured on the main page of the German Wikipedia, or because of a real-world news event (e.g. earthquakes). The result is that an increased exposure predictably leads to increase of both consumption and curation on neighbouring pages, as measured in terms of page requests (for consumption) and edits (for curation), though the author reports that content generation is small in absolute terms.

New papers on the use of Wikipedia in education, by practitioners

In a Portuguese-language conference paper, Brazilian Wikipedian and professor Juliana Bastos Marques “presents an experience with critical reading and edition of Portuguese Wikipedia articles in the university, in extension activities, conducted at the Federal University of Rio de Janeiro State (Unirio), in 2012”, according to the English abstract. In an essay for the sociology journal *Contexts*,^[11] Wikipedian and sociologist (and contributor to other parts of this research newsletter) Piotr Konieczny, who has also made Wikipedia the subject of his own teaching, discusses the benefits of Wikipedia use in academia, citing the view that “a primary reason for academic reservations about Wikipedia is [a] philosophy of knowledge based on the control and management of intellectual capital”.

“World’s largest study on Wikipedia: Better than its reputation”

This is the title of the Helsinki Times' English-language summary of a study of the Finnish Wikipedia's reliability, carried out by journalists and published in the Finnish newspaper *Helsingin Sanomat*.^[12] Participating researcher Arto Lanamäki explained on the Wiki-research-I mailing list that the superlative referred to the

fact that the study had “the biggest sample of articles (134) of all studies that have assessed Wikipedia content quality/credibility.” Not too dissimilar to the approach of the landmark *Nature* study from 2005, the authors recruited “an university-level researcher with knowledge on the subject matter to be an evaluator” for each article in their sample. As summarized by the Helsinki Times, the result was that “the Finnish Wikipedia is largely error-free. The lack of errors is the area in which Wikipedia clearly got its best score. ... No less than 70 per cent of the articles were judged to be good (4) or excellent (5) with respect to lack of errors. According to the indicative evaluation scale a four means that the article has only 'scattered small errors, no big ones'.” (See also earlier coverage of studies that systematically evaluate the reliability of Wikipedia articles: "Pilot study about Wikipedia's quality compared to other encyclopedias", "90% of Wikipedia articles have 'equivalent or better quality than their Britannica counterparts' in blind expert review")

3.0.9 References

- [1] Morgan, J. T.; Gilbert, M.; McDonald, D. W.; Zachry, M. (2014). “Editing beyond articles: diversity & dynamics of teamwork in open collaborations”. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. p. 550. doi:10.1145/2531602.2531654. ISBN 9781450325400.
- [2] André, P.; Kittur, A.; Dow, S. P. (2014). “Crowd synthesis: extracting categories and clusters from complex data”. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. p. 989. doi:10.1145/2531602.2531653. ISBN 9781450325400.
- [3] Mejova, Y.; Garimella, V. R. K.; Weber, I.; Dougal, M. C. (2014). “Giving is caring: understanding donation behavior through email”. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. p. 1297. doi:10.1145/2531602.2531611. ISBN 9781450325400.
- [4] Nakamura, Akira; Yu Suzuki, and Yoshiharu Ishikawa (November 17, 2013). “Clustering Editors of Wikipedia by Editor's Biases” (PDF).
- [5] Xiao, Lu; Nicole Askin (2014). “What influences online deliberation? A wikipedia study”. *Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23004. ISSN 2330-1643.
- [6] Max Bäckström: The conceptualisation of FEMININITY on English Wikipedia <http://www.lunduniversity.lu.se/o.o.i.s?id=24923&postid=4251474>
- [7] Reutner, Ursula (2013-12-20). “Wikipedia und der Wandel der Wissenschaftssprache”. *Romanistik in Geschichte und Gegenwart* **19** (2): 231–249.

- [8] Forte, A., Andalibi, N., Park, T., and Willever-Farr, H. (2014) Designing Information Savvy Societies: An Introduction to Assessability. In: Proceedings of CHI 2014 <http://www.andreaforte.net/ForteCHI14Assessability.pdf>
- [9] Moreira, Carlos Eduardo M.; Viviane P. Moreira (2013-12-09). "Finding Missing Cross-Language Links in Wikipedia". *Journal of Information and Data Management* **4** (3): 251. ISSN 2178-7107.
- [10] Kummer, Michael (2013). "Spillovers in Networks of User Generated Content – Evidence from 23 Natural Experiments on Wikipedia" (PDF). *ZEW Discussion paper no. 13-098*.
- [11] Konieczny, Piotr (2014-02-01). "Rethinking Wikipedia for the Classroom". *Contexts* **13** (1): 80–83. doi:10.1177/1536504214522017. ISSN 1536-5042.
- [12] Koistinen, Olavi (2013-11-30). "HS selvitti: Näin luotettava Wikipedia on". *HS.fi*.



Supplementary references:

- [1] CSCW '14 website
- [2] Wikimedia Research & Data showcase - February 2014
- [3] e.g. Anderka, M., & Stein, B. (2012, April). A breakdown of quality flaws in Wikipedia. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality (pp. 11-18). ACM. (cf. review: "One in four of articles tagged as flawed, most often for verifiability issues")
- [4] e.g. Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, *27*(4), 71-98.
- [5] Liu, J., & Ram, S. (2009, December). Who does what: Collaboration patterns in the wikipedia and their impact on data quality. In *19th Workshop on Information Technologies and Systems* (pp. 175-180)
- [6] Hecht, Brent; Gergle, Darren (2010). *The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context* (PDF). ACM CHI Conference on Human Factors in Computing Systems. pp. 291–300.

Wikimedia Research Newsletter

Vol: 4 • Issue: 2 • February 2014

This newsletter is brought to you by the Wikimedia Research Committee and The Signpost

Subscribe:  [Email](#)  • [archives] [signpost edition]
[contribute] [research index]

Chapter 4

Issue 4(3): March 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 3 • March 2014 [\[contribute\]](#) [\[archives\]](#) 

Wikipedians' "encyclopedic identity" dominates even in Kosovo debates; analysis of "In the news" discussions; user hierarchy mapped

With contributions by: Federico Leva, Scott Hale, Kim Osman, Jonathan Morgan, Piotr Konieczny, Niklas Laxström, Tilman Bayer and James Heilman

4.0.10 Cross-language study of conflict on Wikipedia

Have you wondered about differences in the articles on [Crimea](#) in the Russian, Ukrainian, and English versions of Wikipedia? A newly published article entitled "Lost in Translation: Contexts, Computing, Disputing on Wikipedia"^[1] doesn't address [Crimea](#), but nonetheless offers insight into the editing of contentious articles in multiple language editions through a heavy qualitative examination of Wikipedia articles about [Kosovo](#) in the Serbian, Croatian, and English editions.

The authors, Pasko Bilic and Luka Bulian from the University of Zagreb, found the main drivers of conflict and consensus were different group identities in relation to the topic ([Kosovo](#)) and to Wikipedia in general. Happily, the authors found the dominant identity among users in all three editions was the "encyclopedic identity," which closely mirrored the rules and policies of Wikipedia (e.g., [NPOV](#)) even if the users didn't cite such policies explicitly. (This echoes the result of a similar study regarding political identities of US editors, see previous coverage: "Being [Wikipedia](#) is more impor-

tant than the political affiliation".) Other identities were based largely on language and territorial identity. These identities, however, did not sort cleanly into the different language editions: "language and territory [did] not produce coherent and homogeneous wiki communities in any of the language editions."

The English Wikipedia was seen by many users as providing greater visibility and thus "seem[ed] to offer a forum for both Pro-Serbian and Pro-Albanian viewpoints making it difficult to negotiate a middle path between all of the existing identities and viewpoints." The Arbitration Committee, present in the English edition but not in the Serbian or Croatian editions, may have helped prevent even greater conflict. Enforcement of its decisions seemed generally to lead to greater caution in the edition process.

In line with previous work showing some users move between language editions, the authors found a significant amount of coordination work between the language editions. One central focus centered around whether other editions would follow the English edition in breaking the article into two separate articles ([Republic of Kosovo](#) and [Autonomous Province of Kosovo and Metohija](#)).

4.0.11 The social construction of knowledge on English Wikipedia

review by *Kim Osman*

Another paper by Bilic, published in *New Media & Society*^[2] looks at the logic behind networked societies and the myth perpetuated by media institutions that there is a center of the social world (as opposed to distributed nodes). The paper goes on to investigate the social processes that contribute to the creation of "mediated centers", by analyzing the talk pages of English Wikipedia's *In The News* (ITN) section.

Undertaking an ethnographic content analysis of ITN talk pages from 2004–2012, Bilic found three issues that were disputed among Wikipedians in their efforts to construct a necessarily temporal section of the encyclopedia. First, that editors differentiate between mass media and Wikipedia as a digital encyclopedia, however what con-

stitutes the border between the two is often contested. Second, there was debate between inclusionists and deletionists regarding the criteria for stories making the ITN section. Third, conflict and discussion occurred regarding English Wikipedia's relevance to a global audience.

The paper provides a good insight into how editors construct the ITN section and how it is positioned on the "thin line between mass media agenda and digital encyclopedia." It would be interesting to see further research on the tensions between the Wikipedia policies mentioned in the paper (e.g. **WP:NOTNEWS**, **NPOV**) and mainstream media trends in light of other studies about Wikipedia's approach to breaking news coverage.

4.0.12 User hierarchy map: Building Wikipedia's Org Chart

If you were to make an org chart of English Wikipedia, what would it look like? A recent study^[3] presented at the 2014 *European Conference on Information Systems* examines whether the organizational hierarchy of Wikipedia is as flat and egalitarian as previous research and popular media have claimed in the past. The researchers point out that the degree to which Wikipedia's actual governance model (and those of other peer production communities) reflect egalitarian principles has seldom been comprehensively examined. Furthermore, a growing body of research has shown that Wikipedia has become increasingly bureaucratic along many dimensions, often in response to new community needs. This suggests that Wikipedia has grown more hierarchical, and less flat, over time.

The researchers develop a taxonomy based on technical **user rights** and the quality assurance, coordination, and conflict resolution tasks commonly associated with those user rights. They use exploratory factor analysis, least square analysis, and qualitative examination of the user right description pages to distill 19 user rights down to 8 social roles. They assemble these roles into a hierarchy according to their *Scope*, *Granting*, *Access*, and *Promotion* relationships. For example, in this hierarchy, editors in the *Security Force* role (**checkusers** and **oversighters**) have more power than administrators (**sysops** and **bureaucrats**) because being a **sysop** is an informal prerequisite for **checkuser** rights, and because **oversighters** can use the **RevisionDelete** extension in **suppressor** mode, blocking access to the content from administrators.

The paper does an excellent job of distilling the complex matrix of technologically mediated power relationships within and across Wikimedia wikis into a relatively simple organizational chart (presented on manuscript page 11). However, other mappings are certainly possible. For example, this analysis excludes the role of bots (and therefore, bot wranglers) within the role ecology. It also does not address the soft power that well-respected veteran community members may wield in some situations.

4.0.13 Briefly

Extracting machine-readable data from Wiktionary

Yet another research group recognised **Wiktionary** as a source of «valuable lexical information» and explored conversion of its full content to a machine-readable format, **LMF**.^[4] The **UBY** tools were used as base, but results are not released, probably being in the works (only English, French and German Wiktionaries are mentioned), and seem unaware of **DBpedia's Wiktionary RDF extraction**. Authors find a big obstacle in seemingly innocuous **context labels** of the kind "archaic term": this **diachronicity** would force to split such definitions to separate lexicons by age. Instead, they believe it wouldn't be hard to map all the formats and tags used by the various Wiktionary editions and unify them, apparently, in a single lexicon. If delivered (and open-sourced), such a map could help the perennial discussion on how to unify Wiktionary data, recently revived by the **Wikidata plans**.

Wikipedia as a source of proper names in various languages

Another group^[5] managed to automatically extract proper names mentioned in articles of Wikipedias in 18 European languages, collating the different transliterations and attributing certain properties like "given name" and "family name" (similar to what **Wikidata** does, but without using interwiki links). As in the previous work, the conclusion is that **LMF** is suitable for storing such information, with an extension of the format. The impression is that **LMF's** viability is being tested in "real life" to refine said theoretical standard, an effort parallel to **Wikidata's** process of organic growth by trial and error.

“Wikipedia and Machine Translation: killing two birds with one stone”

This^[6] is a case study about machine aided translation from one language to another. In this case, the researchers made volunteers translate 100 short Computer Science articles from Spanish to Basque Wikipedia, totalling to 50 000 words. They used a rule based machine translation system called **Matxin**. Volunteers corrected the machine translation output using **OmegaT**. The machine translation system was adapted by using a collection of **Mozilla** translations.

Following a long established **Apertium** practice, the human corrections were used as source for a tool to make them automatically. They claim 10% increase in accuracy with this tool, but do not report the baseline or corpus for which it was measured. Additionally: they translated **wikilinks** using **Wikidata**; they noted that markup complicated things; even a not very good machine translation output was still useful for volunteer translators.

“Knowledge Construction in Wikipedia: A Systemic-Constructivist Analysis”

In this study^[7] of knowledge construction on Wikipedia, the authors focus on the importance of the social system and social structure in influencing the actions of individuals (Wikipedia editors). They analyze the edit history of the German Wikipedia article on Fukushima-Daiichi nuclear power plant, arguing that it is a case study of “a regularly occurring situation: the development of new knowledge in a large-scale social setting based on inconsistent information under uncertainty.” The author provide an interesting literature review of what they term a “systemic-constructivist” approach, then discuss the evolution of the Wikipedia article through about 1,200 edits, noting the importance of Wikipedia policies, which were often quoted by the editors. The authors also conducted a survey among the editors of the article to obtain additional information. The authors also asked independent experts to review the article; this review concluded that the German Wikipedia article is of high quality. They note that the experts identified some errors, although unfortunately they do not provide details specific enough for the community to address them. They conclude that the Wikipedia editors were not experts in the field of nuclear power plants, yet were able to produce an article that earned favorable reviews from such experts; this, according to the authors, can be explained through the “systemic-constructivist” approach as validating the importance of the social system and structure of Wikipedia, which guided the amateur editors into producing an expert-level product.

Younger librarians more supportive of Wikipedia

A survey^[8] of information literacy librarians shows that they provide little Wikipedia instruction, with about 40% of respondents answering that their schools provide no instruction on Wikipedia, and 80%, that they hold no dedicated workshops. Still, the remaining group – 60% which do provide some instruction, and 20% who hold dedicated workshops, suggest that the picture is not so dire, and in fact illuminates an interesting opportunity for reaching out with regards to the Wikipedia Education Programs, which do not usually focus on the libraries instructional programs. Only 3% of respondents indicated that they have students actually edit Wikipedia, and one cited story, about “making edits to lower the quality of an article” and “getting a student blocked”, raises a specter of similar incidents in the past (see e.g. previous *Signpost* coverage of a prominent case at George Mason University), as well as a question of ethics in education with regards to purposefully engaging in vandalism for educational purposes. Unsurprisingly, there was also a negative correlation between librarian’s age and views on Wikipedia. Although overall majority of respondents were supportive of the idea that librarians need to educate students in digital literacy skills, they were nonethe-

less opposed to linking Wikipedia from the pages of their institutions.

“Preparing and publishing Wikipedia articles are a good tool to train project management, teamwork and peer reviewed publishing processes in life sciences”

This is the conclusion in the title of a recently published paper from the 2012 “Improving University Teaching” conference^[9] by two zoologists from the University of Innsbruck.

“Networked Grounded Theory” analysis of views on the use of Wikipedia in education

A report paper^[10] describes how a Greek PhD thesis studied the use of Wikipedia in Education using the network visualization software Gephi. Empirical data was gathered “from interviews and focus group discussions with students and teachers participating in Wikipedia assignments, from online blog posts expressing students’, instructors’, and Wikipedians’ reflections on the topic and from Wikipedia’s community discussion pages” and analyzed in a grounded theory approach (classifying text statements into codes such as “Need for Wiki Literate Professors”, “Valuable Content Added”, “You Are Not Listening & Respecting Us” or “Aggressive Community Editors”). Gephi was used to create a visualization grouping these codes (opinions), and grouping them into “communities”. Eventually, the author arrived at “Community Resistance, Organization of Intervention, Community Benefit, Educational Benefit, and Acculturation Stress [as] the conceptual blocks of theory for interpreting the utilization of a virtual community in education as an acculturation process.”

“Risk factors and control of hospital acquired infections: a comparison between Wikipedia and scientific literature”

This is the title of a paper^[11] published in 2013 which analysed Wikipedia content from November of 2010. They looked at 15 articles pertaining to hospital acquired infections (HAIs) of which 8 were B class and the rest were lower. Some of the articles were in this reviewer’s opinion only tangentially related, such as necktie. They looked at how well Wikipedia’s content in 2010 matched the National Institute of Clinical Excellence (NICE) topic on HAIs. NICE writes how to-guides for physicians, while Wikipedians are writing an encyclopedia. The conclusions was thus not surprising that Wikipedia is not a good “how to guide” regarding HAIs (as one editor observed in a discussion about the paper at WikiProject Medicine: “We are criticised for (somewhere) mentioning or recommending signs reminding about hand-

washing routines, ... and for not giving all sorts of detailed guidelines about procedures for the use of catheters and the like by medical staff"). Still, a number of specific errors were also found. Most had already been fixed and this reviewer has corrected the last few.

How a country's broadband connectivity and Wikipedia coverage are related

In 2011, the Oxford Internet Institute began a project to study the online representation of the Arab world online, via Wikipedia. The first peer-reviewed paper from this research became available in preprint form^[12] at the beginning of 2014. As previously observed by these and other researchers, the density of geotagged Wikipedia is highly uneven, and a part of the paper studies its relationship to a country's population, to the number of broadband internet connections in a geographic area, and to Wikipedia's country-level usage statistics over time. Among other things, the authors find that "over three quarters of the variation in geotagged articles was explained by the population of the country, the number of fixed broadband connections and the number of edits emanating from that country." Curiously, the relationship between internet connectivity and Wikipedia coverage was not linear: "those countries with the least and most broadband have more articles than expected, whereas those countries in the middle of the distribution have fewer articles than expected."



4.0.14 References

- [1] Bilic, Pasko and Bulian, Luka (2014). "Lost in Translation: Contexts, Computing, Disputing on Wikipedia". *iConference 2014*.
- [2] Bilic, Pasko (2014). "'Searching for a centre that holds' in the network society: Social construction of knowledge on, and with, English Wikipedia" (PDF). *New Media & Society*. doi:10.1177/1461444814522953. ISSN 1461-4448.
- [3] Arazy, Ofer; Oded Nov, Felipe Ortega (2014). *The [Wikipedia] world is not flat: On the organizational structure of online production communities* (PDF). Twenty Second European Conference on Information Systems.
- [4] *Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data*. Chapter written by Thierry DECLERCK, Pirsoka LENDVAI and Karlheinz MÖRTH. <http://onlinelibrary.wiley.com/doi/10.1002/9781118712696.ch12/summary>
- [5] *Global Atlas: Proper Nouns, From Wikipedia to LMF*. Chapter written by Gil FRANCOPOULO, Frédéric MARCOUL, David CAUSSE and Grégory PIPARO.
- [6] Alegria I., Cabezon U., Fernandez de Betoño U., Labaka G., Mayor A., Sarasola K. and Zubiaga A.: Wikipedia and Machine Translation: killing two birds with one stone. Workshop on 'Free/open-source language resources for the machine translation of less-resourced languages' at LREC 2014. <https://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1395737124>
- [7] Oeberst, Aileen; Iassen Halatchliyski, Joachim Kimmerle, Ulrike Cress (2014-02-21). "Knowledge Construction in Wikipedia: A Systemic-Constructivist Analysis". *Journal of the Learning Sciences*. doi:10.1080/10508406.2014.888352. ISSN 1050-8406.
- [8] Zlatos, Christy (2014-03-12). "Still Not Ready for Prime Time: Academic Librarian Attitudes towards Wikipedia in a Networked Age".
- [9] Schwerte, Thorsten; Stefan Stolz. "Preparing and publishing Wikipedia articles are a good tool to train project management, teamwork and peer reviewed publishing processes in life sciences". *2012 Proceedings*. Improving University Teaching. The University of British Columbia.
- [10] Alexios V. Brailas: Networked Grounded Theory. The Qualitative Report 2014 Volume 19, How To Article 3, 1–16 <http://www.nova.edu/ssss/QR/QR19/brailas3.pdf>
- [11] Maggi, Elisa; Luca Magistrelli, Marco Zavattaro, Marta Beggiano, Fabio Maiello, Cristina Naturale, Margherita Ragliani, Marco Varalda, Maria Sofia Viola, Diego Concina, Elias Allara, Fabrizio Faggiano, Avogadro Wikipedia and HAI Group (2013). "Risk factors and control of hospital acquired infections: a comparison between Wikipedia and scientific literature". *Epidemiology, Biostatistics and Public Health* **10** (1). ISSN 2282-0930.
- [12] Graham, Mark; Bernie Hogan, Ralph K. Straumann, Ahmed Medhat (2014-01-21). *Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty*. Rochester, NY: Social Science Research Network. to appear in Annals of the Association of American Geographers

Wikimedia Research Newsletter

Vol: 4 • Issue: 3 • March 2014

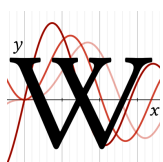
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  Email  • [archives] [signpost edition] [contribute] [research index]

Chapter 5

Issue 4(4): April 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 4 • April 2014 [contribute] [archives] 

Wikipedia predicts flu more accurately than Google; 43% of academics have edited Wikipedia

With contributions by: Piotr Konieczny, Giovanni Luca Ciampaglia and Tilman Bayer

5.0.15 Wikipedia Usage Estimates Prevalence of Influenza-Like Illness

Researchers from [Harvard Medical School](#) have tested the possibility of predicting the number of seasonal influenza-like illness (ILI) in the U.S. using data about the traffic to a selected number of Wikipedia entries related to influenza.^[1]

They compared their models against the prediction of [Google Flu Trends \(GFT\)](#), one of the earliest and most famous web-based tools for predicting the evolution of seasonal influenza disease patterns. [Gold standard](#) for comparison were the public data released by the [Center for Disease Control \(CDC\)](#). The accuracy of GFT is increasingly under question by several authors, culminating in a recent [Science](#) commentary piece about the promises and perils of Big Data for prediction of real-world phenomena. The authors start from this observation and submit that Wikipedia searches may be less subject to the biases that affected GFT, and test this hypothesis in the present work. They find that their model is more accurate than GFT, and was able to predict the peak week of the influenza season more often. Another undoubted advantage of Wikipedia compared to GFT, the authors argue, is its public availability, which makes the present model

open to public scrutiny.

5.0.16 Survey of academics' view on Wikipedia and open-access publishing

A study titled “Academic opinions of Wikipedia and open-access publishing”^[2] examined academics' awareness of and attitudes towards Wikipedia and open-access journals for academic publishing through a survey of 120 academics carried out in late 2011 and early 2012. The study comes from the same authors who published a similar paper in 2012, reviewed [here](#), which suffered from a major basic fallacy: Wikipedia is not the place to publish original research academic work. The authors, unfortunately, seem to ignore [no original research policy](#) when they write: “There are in general three models in the current movement towards open-access academic publishing: pushing traditional journals towards open access by changing policies; creating open-access journals; and using existing online open-access venue Wikipedia” and “we surveyed academics to understand their perspectives on using Wikipedia for academic publishing in comparison with open-access journals”. In the final discussion segment, the authors do acknowledge the existence of the OR policy, where they suggest that certain types or academic papers (reviews) are similar enough to Wikipedia articles that integration of such articles into Wikipedia could be feasible. The authors do provide a valuable literature review noting prior works which analyze the peer-review system in Wikipedia, perceptions of Wikipedia in academia, and related issues (through said review is partially split between the introduction and discussion section).

The study provides some interesting findings regarding academics' view of the benefits of Wikipedia-style peer review and publishing. Most respondents (77 percent) reported reading Wikipedia, and a rather high number (43 percent) reported having made at least one edit, with 15 percent having written an article. Interestingly, as many as four respondents stated that they were “credited for time spent reviewing Wikipedia articles related to their academic careers” in their professional workplaces. The

more experience one had with Wikipedia, the more likely one would see advantages in the wiki publishing model. Most common advantages listed were cost reductions (40 percent), timely review (19 percent), post-publication corrections (52 percent), making articles available before validation (27 percent) and reaching a wider audience (8 percent). Disadvantages included questionable stability (86 percent), absence of integration with libraries and scholarly search engines (55 percent), lower quality (43 percent), less credibility (57 percent), less academic acceptance (78 percent) and less impact on academia (56 percent).

54 percent of respondents were aware that Wikipedia had a peer-review process and about third of these considered it to be less rigorous than that of scholarly journals; none of the respondents demonstrated any significant experience with the specifics of how Wikipedia articles are reviewed, suggesting that their involvement with the Wikipedia is rather limited. 75% of the survey respondents did not feel comfortable having others edit their papers-in-progress, and over 25% expressed concern about the lack of control over changes made post-publications. Majority of respondents did not also feel comfortable with their work being reviewed by Wikipedians, with the most common concern being unknown qualifications of Wikipedia editors and reviewers.

Perhaps of most value to the Wikipedia community is the analysis of suggestions made by the respondents with regards to making Wikipedia more accepted at the universities. Here, the most common suggestion was “making the promoted peer-reviewed articles searchable from university libraries” and in general, making it more easy to find and identify high quality articles (some functionality as displaying the quality assessment of an article in mainspace already exists in MediaWiki but is implemented as opt-in feature only).

The authors conclude that the academic researchers’ increased familiarity with either open access publishing or wiki publishing is associated with increased comfort with these models; and the academic researchers’ attitudes towards these models are associated with their familiarity, academic environment and professional status. Overall, this study seems like a major improvement over the authors’ 2012 paper, and a valuable paper addressing the topics of the place of Wikipedia in the open publishing movement and the relationship between Wikipedia and academia.

5.0.17 Briefly

Wikipedia use driven by news media or replacing news media?

In a series of blog posts^{[3][4][5]} Oxford Internet Institute researchers Taha Yasseri and Jonathan Bright examined pageview data from before, during and after the

2009 European Parliament election on different language Wikipedias (mostly corresponding to different European countries where the election took place). They found evidence both for the theory that Wikipedia readership is driven by media coverage (people turning to Wikipedia for background information on what they see in the news) and for the theory that Wikipedia acts as “media replacement” (people looking online for e.g. election results instead of getting that information from news media).

New Python library for researchers

Wikimedia Foundation researcher Aaron Halfaker published a collection of software tools “for extracting and processing data from MediaWiki installations, slave databases and xml dumps.”

“Do Famous People Live Longer?” Yes for academics, no for artists and athletes

Four researchers from Ben-Gurion University of the Negev examined^[6] 7756 biographical Wikipedia articles about people who had died between 2009 and 2011 for gender, occupation and age at death. 84% of the article subjects were male, “and the mean age of death was lower for males than females (76.31 vs. 78.50 years). Younger ages of death were evident among sports players and performing artists (73.04) and creative workers (74.68). Older deaths were seen in professionals and academics (82.63).” Two of the authors also published another preprint titled “Wikiometrics: A Wikipedia Based Ranking System”^[7], applying it to universities and academic journals in particular. The resulting rankings correlate strongly with some established metrics like impact factors.

5.0.18 Other recent publications

A list of other recent publications that could not be covered in time for this issue - contributions are always welcome for reviewing or summarizing newly published research.

- “Behavioral Aspects in the Interaction Between Wikipedia and its Users”^[8] (see also our review of an earlier paper that the two authors published with others in 2012: “Science eight times more popular on the Spanish Wikipedia than on the English Wikipedia?”)
- “Bots vs. Wikipedians, Anons vs. Logged-Ins”^[9] (poster at the WWW 2014 conference)
- “Telling Breaking News Stories from Wikipedia with Social Multimedia: A Case Study of the 2014 Winter Olympics”^[10]

- “A classifier to determine which Wikipedia biographies will be accepted”^[11] - according to the abstract, it relies on “indicators [that] do not refer to the content itself, but to meta-content features (such as the number of categories that the biography is associated with) and to author-based features (such as if it is a first-time author)”.
 - “What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data”^[12]
 - “Counter narratives and controversial crimes: The Wikipedia article for the ‘Murder of Meredith Kercher’”^[13] (a linguistic essay examining two different versions of the article each on the English and the Italian Wikipedia. University press release: “Scrutinising the myth of social media ‘neutrality’”)
 - “Assessing the Quality of Thai Wikipedia Articles Using Concept and Statistical Features”^[14]
 - “The Genealogy of Knowledge: Introducing a Tool and Method for Tracing the Social Construction of Knowledge on Wikipedia”^[15]
 - “Wikipedia As a Tool for Disseminating Knowledge of (Agro)Biodiversity”^[16]
 - “Complementary and Alternative Medicine on Wikipedia: Opportunities for Improvement”^[17]
 - “Revision Graph Extraction in Wikipedia Based on Supergram Decomposition and Sliding Update”^[18] (earlier coverage of related papers by the same authors: “Revision graph extraction in Wikipedia based on supergram decomposition”, “Unearthing the “actual” revision history of a Wikipedia article”)
 - “Detecting Controversial Articles in Wikipedia”^[19] (as an exercise in an undergraduate course on graph theory)
- [5] Taha Yasseri, Jonathan Bright. “Media effect or media replacement?”. *Can social data be used to predict elections?*.
- [6] Nir Ofek, Lior Rokach, Armin Shmilovici, Gilad Katz: Do Famous People Live Longer? A Wikipedia Analysis. ResearchGate, January 2014. PDF
- [7] Lior Rokach, Gilad Katz: Wikiometrics: A Wikipedia Based Ranking System. ResearchGate, January 2014. PDF
- [8] Reinoso, Antonio J.; Juan Ortega-Valiente (2014-01-01). “Behavioral Aspects in the Interaction Between Wikipedia and its Users”. In Cristian Lai, Alessandro Giuliani, Giovanni Semeraro (eds.). *Distributed Systems and Applications of Information Filtering and Retrieval*. Studies in Computational Intelligence. Springer Berlin Heidelberg. pp. 135–149. ISBN 978-3-642-40621-8. :10.1007/978-3-642-40621-8_8
- [9] Steiner, Thomas (2014-02-03). “Bots vs. Wikipedians, Anons vs. Logged-Ins”. *arXiv:1402.0412 [cs]*.
- [10] Steiner, Thomas (2014-03-17). “Telling Breaking News Stories from Wikipedia with Social Multimedia: A Case Study of the 2014 Winter Olympics”. *arXiv:1403.4289 [cs]*.
- [11] Ofek, Nir; Lior Rokach (2014-05-01). “A classifier to determine which Wikipedia biographies will be accepted”. *Journal of the Association for Information Science and Technology*: â€“. doi:10.1002/asi.23199. ISSN 2330-1643.
- [12] Lucie Flekova, Oliver Ferschke, and Iryna Gurevych What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data http://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2014/WWW2014_WikiAFT.pdf Preprint of an article accepted for publication in the proceedings of the 23rd International World Wide Web Conference
- [13] Page, Ruth (2014-02-01). “Counter narratives and controversial crimes: The Wikipedia article for the ‘Murder of Meredith Kercher’”. *Language and Literature* **23** (1): 61–76. doi:10.1177/0963947013510648. ISSN 0963-9470.
- [14] Kanchana Saengthongpattana, Nuanwan Soonthornphisaj: Assessing the Quality of Thai Wikipedia Articles Using Concept and Statistical Features, p. 513 in: *New Perspectives in Information Systems and Technologies*, Volume 1. Editors: Álvaro Rocha, Ana Maria Correia, Felix B Tan, Karl A Stroetmann. ISBN: 978-3-319-05950-1 (Print) 978-3-319-05951-8 (Online)
- [15] Friedrich Chasin, Uri Gal, Kai Riemer: The Genealogy of Knowledge: Introducing a Tool and Method for Tracing the Social Construction of Knowledge on Wikipedia. 24th Australasian Conference on Information Systems, 4-6 Dec 2013, Melbourne
- [16] Signore, Angelo; Francesco Serio, Pietro Santamaria (2014-02-01). “Wikipedia As a Tool for Disseminating Knowledge of (Agro)Biodiversity”. *HortTechnology* **24** (1): 118–126. ISSN 1063-0198.

5.0.19 References



- [1] McIver, David J; John S. Brownstein (2014-04-17). “Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time”. *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1003581.
- [2] Xiao, Lu; Nicole Askin (2014-04-29). “Academic opinions of Wikipedia and open-access publishing”. *Online Information Review* **38** (3). ISSN 1468-4527.
- [3] Taha Yasseri, Jonathan Bright. “The electoral information cycle”. *Can social data be used to predict elections?*.
- [4] Taha Yasseri, Jonathan Bright. “Outliers on the electoral information cycle.”. *Can social data be used to predict elections?*.

- [17] Koo, Malcolm (2014-04-17). “Complementary and Alternative Medicine on Wikipedia: Opportunities for Improvement”. *Evidence-Based Complementary and Alternative Medicine* **2014**. doi:10.1155/2014/105186. ISSN 1741-427X.
- [18] Wu, Jianmin; Mizuho Iwaihara (2014-04-01). “Revision Graph Extraction in Wikipedia Based on Supergram Decomposition and Sliding Update”. *IEICE TRANSACTIONS on Information and Systems*. E97-D (4): 770–778. ISSN 1745-1361.
- [19] Joy Lind, Darren A. Narayan: Detecting Controversial Articles in Wikipedia PDF

Wikimedia Research Newsletter

Vol: 4 • Issue: 4 • April 2014

This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  **Email**  • [archives] [signpost edition]
[contribute] [research index]

Chapter 6

Issue 4(5): May 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 5 • May 2014 [contribute] [archives] 

Overview of research on Wikipedia's readers; predicting which article you will edit next

With contributions by: Piotr Konieczny, Maximilian Klein and Tilman Bayer

6.0.20 “Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership”

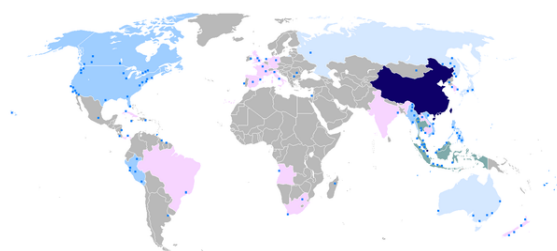
This paper ^[1] is another major literature review of the field of Wikipedia studies, brought forward by the authors whose prior work on this topic, titled “The People’s Encyclopedia Under the Gaze of the Sages”^[supp 1] was reviewed in this research report in 2012 (“A systematic review of the Wikipedia literature”).

This time the authors focus on a fragment of the larger body of works about Wikipedia, analyzing 99 works published up to June 2011 on the theme of “Wikipedia readership” – in other words focusing on the theme “What do we know about people who read Wikipedia”. The overview focuses less on demographic analysis (since little research has been done in that area), and more on perceptions of Wikipedia by surveyed groups of readers. Their findings include, among other things, a conclusion that “Studies have found that articles generally related to entertainment and sexuality top the list, covering over 40% of visits”, and in more serious topics, it is a common source for health and legal information. They also find

that “a very large number of academic in fact have quite positive, if nuanced, perceptions of Wikipedia’s value.” They also observe that the most commonly studied group has been that of students, who offer a convenience sample. The authors finish by identifying a number of contradictory findings and topics in need of further research, and conclude that existing studies have likely overestimated the extent to which Wikipedia’s readers are cautious about the site’s credibility. Finally, the authors offer valuable thoughts in the “implications for the Wikipedia community” section, such suggesting “incorporating one or more of the algorithms for computational estimation of the reliability of Wikipedia articles that have been developed to help address credibility concerns”, similar to the WikiTrust tool.

The authors also published a similar literature review paper summarizing research about the *content* of Wikipedia, which we hope to cover in the next issue of this research report.

6.0.21 Chinese-language time-zones favor Asian pop and IT topics on Wikipedia



Map of the Chinese-speaking world

A paper^[2] presented at the *WWW 2014 Companion Conference* analyzes the readership patterns of the English and Chinese Wikipedias, with a focus on which types of articles are most popular in the English- or Chinese-language time zones. The authors used all Wikipedia pages which existed under the same name in both languages in the period from 1 June 2012 to 14 October 2012 for their study, coding them through the OpenCalais

semantic analysis service with an estimated 2.6% error rate.

The authors find that readers of the English and Chinese Wikipedias from time-zones of high Chinese activity browse different categories of pages. Chinese readers visit English Wikipedia about Asian culture (in particular, Japanese and Korean pop culture) more often, as well as about mobile communications and networking technologies. The authors also find that pages in English are almost ten times as popular as those in Chinese (though their results are not identifying users by nationality directly, rather focusing on time zone analysis).

In this reviewer's opinion, the study suffers from major methodological problems that are serious enough to cast all the findings in doubt. Apparently because the authors were unaware of *interlanguage links* and consider only articles which have the same name (URL) in both the English and Chinese Wikipedias, they find that only 7603 pages were eligible to be analyzed (as they had both an English and Chinese version), however the Chinese Wikipedia in the studied period had approximately half a million articles; and while many don't have English equivalents yet, to expect that less than 2% did seems rather dubious. Similarly, our own WikiProject China estimates that English Wikipedia has almost 50,000 China-related articles. That, given that WikiProject assessments are often underestimating the number of relevant topics, and usually don't cover many core topics, suggests that the study missed a vast majority of articles that exist in both languages. It is further unclear how English- and Chinese-language time-zones were operationalized. The authors do not reveal how, if at all, they controlled for the fact that readers of English Wikipedia can also come from countries where English is not a native language, and that there are hundreds of millions of people outside China who live in the five time zones that span China, which overlap with India, half of Russia, Korea and major parts of Southeast Asia. As such, the findings of that study can be more broadly interpreted as “readership patterns of English and Chinese Wikipedia in Asia and the world, regarding a small subset of pages that exist on both English and Chinese Wikipedia.”

6.0.22 “Bipartite editing prediction in Wikipedia”

Reviewed by Maximilianklein (talk)

Bipartite Editing Prediction in Wikipedia^[3] is a paper wherein the authors aim to solve what they call the “link prediction problem”. Essentially they aim to answer “which editors will edit which articles in the future.” They claim the social utility of this is to suggest articles to edit to users. So in some ways this is a similar function to SuggestBot, but using different techniques.

Their approach here is to use a bipartite network mod-

elling. A bipartite network is a network with two node-types, here editors and articles. Using bipartite network modelling is becoming increasingly trendy, like Jesus (2009)^[supp 2] and Klein (2014).^[supp 3]

Explaining their method, the researchers outline their two approaches: “supervised learning” and “community awareness”. In the supervised learning approach the machine learning features used are Association Rule, K-nearest neighbor, and graph partitions. All these features, they state, can be inferred directly from the bipartite network. In the community awareness approach, the Stanford Network Analysis Project tool is used to cut the network into co-editor sets, and then go on to inspect what they call indirect features which are sum of neighbors, Jaccard coefficient, preferred attachment, and Adamic-Adar score.

The authors proceed to give a table of their results, and highlight their highest achieving precision, and recall statistics which are moderate and contained in the interval [.6, .8]. Thereafter a short non-interpretive one-paragraph discussion concludes the paper saying that these results might be useful. Unfortunately they are not of much use, since while they declare their sample size of 460,000 editor–article pairs from a category in a Wikipedia dump, they don't specify which category, or even which Wikipedia they are working on.

This machine learning paper lacks sufficient context or interpretation to be immediately valuable, despite the fact that they may be able to predict with close to 80% F-measure which article you might edit next. Therefore the paper is a good example of the extent to use Wikipedia for research without even feigning attempt to make the research useful to the Wikipedia community, or even frame it in that way.

6.0.23 Briefly



A reading room in the University of Pittsburgh's Hillman Library

“Increasing the discoverability of digital collections using Wikipedia: the Pitt experience”

In this paper,^[4] a librarian at the University of Pittsburgh discusses how two undergraduate interns have added over 100 links to library collections to Wikipedia articles, which led to the increase use of the library’s digitized collections. An experienced Wikipedian, Sage Ross, provided help with this project. The two undergrads expanded or created approximately 100 articles, mainly related to the History of Pittsburgh (such as Pittsburgh Courier or Pittsburgh Playhouse), using resources hosted by the university’s libraries as sources or external links. The paper also provides a valuable overview of similar initiatives in the past (some of which have also been covered in this research report, see e.g.: "Using Wikipedia to drive traffic to library collections"). The majority of reviewed examples suggest that linking library resources from Wikipedia pages increases their visibility, and this study reached the same conclusion with regards to their project, which led both the improvement of Wikipedia content and of driving more traffic to the digital resources hosted by the library. This reviewer applauds this project as a model one, though it would benefit from a list of all articles edited by the students (which were not tagged on their talk pages with any expected template, such as {{educational assignment}}).

Korean survey on “Key Factors for Success” of Wikipedia and Q&A site

This paper^[5] compares aspects of Wikipedia and South Korean Naver’s “Naver Knowledge” service (see Knowledge Search), similar to Google Questions and Answers. This is a topic of some interest, as South Korea is praised for being one of the most Internet-integrated societies in the world, while at the same time the Korean Wikipedia currently holding the rank of 23rd largest, is less developed than those of a number of smaller countries less commonly seen as Internet powers (consider List of Wikipedias by size). The researchers surveyed 132 Korean Internet users of those services, though they do not make it clear if all members of the sample were in fact registered contributors to both services, instead describing them as “relative active users of the CI [collective intelligence] system”. Unfortunately, parts of the paper, including the survey questions, appear to have been translated using machine translation, and are thus difficult to interpret correctly. Overall, the authors find that there were no significant differences with regards to the respondents views of Naver Knowledge and Wikipedia services. One of the statistically significant results suggest that Korean contributors of collective intelligence services find the Naver Knowledge service easier to use than Wikipedia, though the differences do not appear to be major (73.5% and 60.9% of Korean contributors found Naver Knowledge and Wikipedia easy to work with, respectively). One of the conclusions of the paper is the

importance of making user interfaces as easy as possible, and making it easier for the users to add and edit audio-visual content (though the authors seem not aware of and do not discuss the Visual Editor).

“Citation filtered”

This glossy and infographic-laden report dissects the 963 Persian Wikipedia articles that are blocked in Iran.^[6] The technique used was to programmatically iterate over Wikipedia to see which articles could not be loaded. Categorizing the articles into 10 topics, an analysis of the Iranian Government’s sensitivities are explored. From the Annenberg School of Communication, University of Pennsylvania blog. (*Maximilianklein (talk)*)

“Georeferencing Wikipedia documents using data from social media sources”

This paper^[7] describes several methods to automatically assign geocoordinates to articles on the English Wikipedia, by matching the article text: to hashtags of georeferenced tweets; to tags of georeferenced photos on Flickr; and to the text of other Wikipedia articles that are already georeferenced. The authors report that “using a language model trained using 376K Wikipedia documents, we obtain a median error of 4.17 km, while a model trained using 32M Flickr photos yields a median error of 2.5 km. When combining both models, the median error is further reduced to 2.16 km. Repeating the same experiment with 16M tweets as the only training data results in a median error of 35.81 km”. As one possible application, the authors suggest automatic correction of coordinates for Wikipedia articles where their method predicts a differing location with high confidence. Among their test dataset of 21,839 articles with a geocoordinate located in the United Kingdom, the authors found three such errors, one of which was still uncorrected at the time of their preprint publication (an educational institution in Brussels which had been placed in Cornwall due to a sign error in the longitudinal coordinate). Another interesting byproduct is a visual comparison (figure 5) of the density of geolocated entries from Wikipedia, Twitter and Flickr in Africa (per the datasets used).

6.0.24 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- **"Snuggle: Designing for efficient socialization and ideological critique"**^[8]
- **“Preferences in Wikipedia abstracts: Empirical**

findings and implications for automatic entity summarization”^[9]

- **“Cluster approach to the efficient use of multimedia resources in information warfare in wikipedia”^[10]** (from the abstract: “A new approach to uploading files in Wikimedia is proposed with the aim to enhance the impact of multimedia resources used for information warfare in Wikimedia.”)
- **“From open-source software to Wikipedia: ‘Backgrounding’ trust by collective monitoring and reputation tracking”^[11]** (from the abstract: “It is shown that communities of open-source software—continue to—rely mainly on hierarchy (reserving write-access for higher echelons), which substitutes (the need for) trust. Encyclopedic communities, though, largely avoid this solution. In the particular case of Wikipedia, which is confronted with persistent vandalism, another arrangement has been pioneered instead. Trust (i.e. full write-access) is ‘backgrounded’ by means of a permanent mobilization of Wikipedians to monitor incoming edits. ... Finally it is argued that the Wikipedian monitoring of new edits, especially by its heavy reliance on computational tools, raises a number of moral questions that need to be answered urgently.”)

6.0.25 References

- [1] Okoli, Chitu and Mehdi, Mohamad and Mesgari, Mostafa and Nielsen, Finn Årup and Lanamäki, Arto (2014): Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the American Society for Information Science and Technology*. ISSN 1532-2882 (In Press) PDF
- [2] Tinati, Ramine; Paul Gaskell; Thanassis Tiropanis; Olivier Phillipe; Wendy Hall (2014). “Examining Wikipedia across linguistic and temporal borders”. *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. WWW Companion '14. International World Wide Web Conferences Steering Committee. pp. 445–450.
- [3] CHANG, YANG-JUI; YU-Chuan Tsai; Hung-Yu Kao (May 2014). “Bipartite editing prediction in Wikipedia”. *Journal of Information Science and Engineering* **30** (3): 587-603.
- [4] Galloway, Ed; Cassandra DellaCorte (2014-05-02). “Increasing the discoverability of digital collections using Wikipedia: the Pitt experience”. *Pennsylvania Libraries: Research & Practice* **2** (1): 84–96. doi:10.5195/palrap.2014.60. ISSN 2324-7878.
- [5] Seo-Young Lee, Sang-Ho Lee, “A Comparison Study on the Key Factors for Success of Social Authoring Systems – focusing on Naver KiN and Wikipedia”, *AISS: Advances in Information Sciences and Service Sciences*, Vol. 5, No. 15, pp. 137 ~ 144, 2013,PDF
- [6] egcsblog. “Citation-Filtered”. Retrieved 31 May 2014.
- [7] Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, Christopher B. Jones: Georeferencing Wikipedia documents using data from social media sources. Preprint, accepted for publication in: *ACM Transactions on Information Systems*, Volume 32 Issue 3 PDF
- [8] Halfaker, Aaron; R. Stuart Geiger; Loren Terveen (2014-04-28). “Snuggle: designing for efficient socialization and ideological critique” (PDF). *CHI: Conference on Human Factors in Computing Systems*. doi:10.1145/2556288.2557313.
- [9] Xu, Danyun; Gong Cheng; Yuzhong Qu (March 2014). “Preferences in Wikipedia abstracts: empirical findings and implications for automatic entity summarization”. *Information Processing & Management* **50** (2): 284–296. doi:10.1016/j.ipm.2013.12.001. ISSN 0306-4573.
- [10] Alguliev, R. M.; R. M. Aliguliyev; I. Ya Alekperova (2014-03-01). “Cluster approach to the efficient use of multimedia resources in information warfare in wikipedia”. *Automatic Control and Computer Sciences* **48** (2): 97–108. doi:10.3103/S0146411614020023. ISSN 0146-4116.
- [11] de Laat, Paul B. (2014-04-22). “From open-source software to Wikipedia: ‘backgrounding’ trust by collective monitoring and reputation tracking”. *Ethics and Information Technology*: 1–13. doi:10.1007/s10676-014-9342-9. ISSN 1388-1957.



Supplementary references:

- [1] Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F., & Lanamäki, A. (2012, October 24). The People’s Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia. SSRN Scholarly Paper, Montreal. <http://papers.ssrn.com/abstract=2021326>
- [2] Rut Jesus; Martin Schwartz; Sune Lehmann (2009). “Bipartite networks of Wikipedia’s articles and authors: a meso-level approach” (PDF).
- [3] Klein. “Measuring Editor Collaborativeness With Economic Modelling”.

Wikimedia Research Newsletter

Vol: 4 • Issue: 5 • May 2014

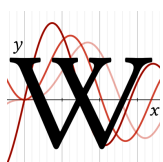
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  [Email](#)  • [archives] [signpost edition] [contribute] [research index]

Chapter 7

Issue 4(6): June 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 6 • June 2014 [contribute] [archives] 

Power users and diversity in WikiProjects; the “network of cultures” in multilingual Wikipedia biographies

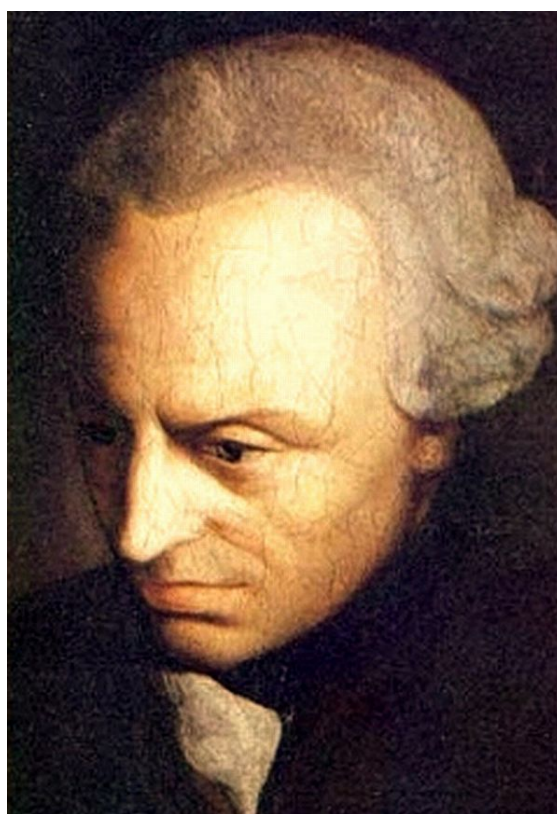
With contributions by: Taha Yasseri, Maximilian Klein, Piotr Konieczny, Kim Osman, and Tilman Bayer

7.0.26 New book: *Global Wikipedia*

An edited volume^[1] by Pnina Fichman and Noriko Hara from Indiana University, Bloomington was released on May 23, 2014, subtitled “International and Cross-cultural Issues in Online Collaboration”. The book description states that “dozens of books about Wikipedia are available, but they all focus on the English Wikipedia and assume an Anglo-Saxon perspective, while disregarding cultural and language variability or multi-cultural collaborative efforts”. The description claims that this is “the first book to address this gap by focusing attention on the global, multilingual, and multicultural aspects of Wikipedia.” The book contains nine chapters authored by 16 Wikipedia researchers (including a chapter authored by the volume editors). Among the topics covered are international and cross-cultural conflict and collaboration, case studies in the Chinese, Finnish, French, and Greek Wikipedias, and Wikipedia gender gaps in different language sites.

7.0.27 “Interactions of cultures and top people of Wikipedia from ranking of 24 language editions”

Review by *Maximilianklein (talk)*



The German philosopher Immanuel Kant, born in today's Russia, is among the small number of cases where the researchers' method of assigning a historical figure to a national culture based on their birth place fails

This research by Eom et al.^[2] is an exploratory data analysis of figures (roughly, “people”) from a mining of date and place of birth and gender in biography articles. Presenting novel ideas based on the infamous Google PageRank algorithm, this paper is a sort of computational history. The methods used are standard – if not a bit dated – compared with more contemporary research using Wikidata. This is a shame because newer techniques

would have allowed the claims of a quantified cultural influence factor to rest on firmer grounds.

Their method is for each of their 24 Wikipedia languages (approximately the top 24 largest ones) to construct the network where nodes are biography articles, and links are intrawiki-links. Then they rank each node by both PageRank and 2DRank. PageRank says your importance is a recursive function of your incoming links, weighted by the page rank of each incoming linker; CheiRank is the same as PageRank, but using outgoing links instead. 2DRank is a mixture of PageRank and CheiRank. Some of the authors have coauthored earlier papers that similarly examined PageRank and CheiRank for biographical and other Wikipedia articles (see our previous coverage: "How Wikipedia's Google matrix differs for politicians and artists" and "Multilingual ranking analysis: Napoleon and Michael Jackson as Wikipedia's 'global heroes'").

However, the input to these algorithms is the weak part. The base set consists of all of the articles that are in a subcategory of Biographies of Living People, Births by Year, or Deaths by Year. Obtaining 1.1 million biography articles, they acknowledge that this isn't a full set because it is based off English Wikipedia, but then make an anecdotal claim that it's only 2% off. However, with the latest Wikidata information we know of at least 2.08 million "people" with Wikipedia articles^[3].

The rest of their method consists of finding the top 100 articles in each of the 24 languages using both PageRank and 2DRank. Then they get birth place, birthdate and gender from DBpedia if available, and if not they look up this information manually. They pigeonhole each article into one of the 24 target cultures based on birth place, and use a "World" category if none applies. Simplifying assumptions are also made during these processes: modern borders are used, and each country is assumed to speak only a single language. So Kant is Russian and all Belgians speak Dutch in this research.

There is an exploratory analysis of these top 100 by geography, time, and gender. The results confirm a long-told story: the biographies that the English Wikipedia knows about are heavily skewed towards being Western/European, modern, and male. They make points of showing local favour, e.g. Hindi has many in their top 100 who are born in India. With regard to history, the authors note that the Arabic Wikipedia is more interested in history than what world growth would suppose. Another measure is defined to look at the localness factor by decade – that is, what percentage of top figures in this decade were born in this language-place? Of course it's Greeks early on, and the US dominating later.

On gender, their results indicate 5.1% or 10.1% by PageRank and 2DRank, respectively, are female of the top 100s, averaged. The authors make mention that maleness does decrease over time as well. This reported figure is more severe than the overlap with any single language, so the authors show some "wisdom of the crowds" effect.

The final analysis tries to quantify cultural influence. A "network of cultures" is made, where nodes are each of the 24 languages-cum-cultures, and the directed, weighted edges are the number of foreigners in their top 100. For instance, in the English Wikipedia's top 100, five people were born in France; so England connects to France with a weight of 5. With this "network of cultures" in hand, they apply the PageRank and 2DRank algorithms to rank each culture. This is a novel approach to making statistical what we all often guess at. Even despite the fact that Jesus is considered Arabic through their simplifications, PageRank turns up English and German as top and runner-up, respectively. Using 2DRank, Greek, French and Russian get more due.

In summary, although this cultural research suffers from biased data, some clever ideas are implemented – particularly the "network of cultures". The implication is that statistical history somewhat corroborates the opinions of manually conducted history.

7.0.28 “Recommending reference materials in context to facilitate editing Wikipedia”

This article^[4] describes *IntelWiki*, a set of MediaWiki tools designed to facilitate new editor's engagement by making research easier. The tool "automatically generates resource recommendations, ranks the references based on the occurrence of salient keywords, and allows users to interact with the recommended references within the Wikipedia editor." The researchers find that volunteers using this tool were more productive, contributing more high-quality text. The studied group was composed of 16 editors with no Wikipedia editing experience, who completed two editing tasks in a sandbox wiki, one using a mockup Wikipedia editing interface and Google search engine, and using the IntelWiki interface and reference search engine. The author's reference suggestion tool seems valuable, unfortunately this reviewer was unable to locate any proof that the developer engaged the Wikipedia community, or made his code or the tool publicly available for further testing. The research and the thesis does not discuss the differences between their MediaWiki clone and Wikipedia in any significant details. Based on the limited description, the study's overall conclusions may not be reliable, since the mockup Wikipedia interface used for the comparison seems to be a default MediaWiki clone, lacking many Wikipedia-specific tools; therefore the theme of comparing *IntelWiki* to *Wikipedia* is a bit misleading.

While the study is interesting, it is disappointing that the main purpose appears to be completing a thesis,^[5] with little thought to actually improving Wikipedia (by developing public tools and/or releasing open code). (See also: related webpage, YouTube video)

7.0.29 “What do Chinese-language microblog users do with Baidu Baike and Chinese Wikipedia?”

This paper ^[6] (accepted for presentation at OpenSym 2014, and subtitled “A case study of information engagement”) explores the use of the Chinese Wikipedia and Baidu Baike encyclopedia by Chinese microblog (Twitter, Sina Weibo) users through qualitative and quantitative analyses of Chinese microblog postings. Both encyclopedias are often cited by microblog users, and are very popular in China to the extent that the words “wiki” and “baidu” have become verbs meaning to look up content on the respective websites, analogous to “to google” in English.

One of the study’s major focuses is the impact of Internet censorship in China; particularly since Wikipedia is not censored – but access to it, and its discussion in most Chinese websites may be. Baidu Baike is both censored and more likely to host copyright violating content. Despite Baidu Baike’s copyright violating content, many users still prefer the uncensored and more reliable Chinese Wikipedia, though they can become frustrated by not being able to access it due to censorship. Whether some Wikipedia content is censored or not is seen by some as a measure of the topic’s political sensitivity. The author suggests that a distinguishing characteristic can be observed between groups that prefer one encyclopedia over the other, but does not discuss this in detail, suggesting a very interesting research avenue.

7.0.30 Content or people? Achieving critical mass to promote growth in WikiProjects

Review by Kimaus

In a recent paper^[7], Jacob Solomon and Rick Wash investigate the question of sustainability in online communities by analysing trends in the growth of WikiProjects. Solomon and Wash track revisions and membership in over one thousand WikiProjects over a period of five years to examine how the concept of a *critical mass* can influence a community’s development. The key question being, as the title of the paper states: “Critical mass of what?” Is it achieving a certain number of contributions or a certain number of members that will ensure the future sustainability of an online group?

Using *critical mass* theory, which describes groups as having an accelerating, linear or decelerating production function, the authors modelled a growth curve for each community. They found that the majority of WikiProjects had an accelerating growth regarding the number of revisions, however a decelerating growth in accruing members which suggests that existing editors are increasing individual contributions to the projects. In further ex-

amining this trend Solomon and Wash focus on the early years of projects’ existence to determine whether amassing content or editors in this formative period influences future production functions.

Their modelling shows that a greater number and diversity of editors within a project positively affects the number of revisions accumulated after five years (where diversity is calculated through membership in other WikiProjects). Interestingly, the modelling showed contributions by infrequent participants helped a project grow, but this can be offset by “overparticipation from a project’s power users.” They attribute this to members’ feeling that they can make a difference to projects that have diverse and sparse contributions. They do note, however, that increased contributions from power users may simply be an attempt to keep a project afloat, and that this effort is ultimately futile in certain cases. In sum, the authors find that it is a critical mass of people (who hold a variety of skills and knowledge) contributing small amounts in the early stages that positively affects a project’s growth and future sustainability.



A cinema audience, possibly containing Wikipedia readers

7.0.31 “Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity”

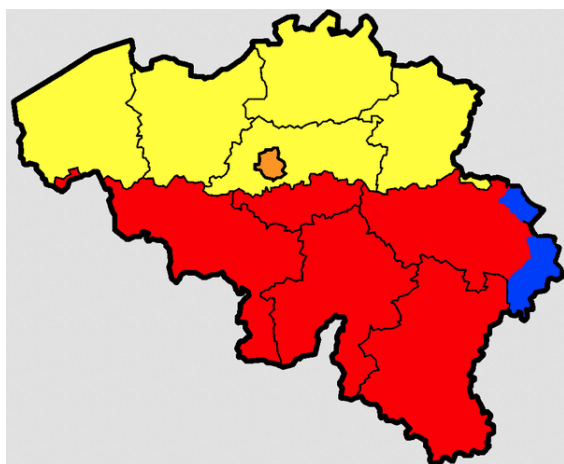
In a paper^[8] presented at the ChASM Workshop of WebSci’14, Bloomington, Indiana, this month, de Silva and Compton, have generalised a method, previously introduced by Mestyán, Yasserli, and Kertész (see the [newsletter review](#)) to predict the box office revenues of movies based on the Wikipedia edits and page-view counts. Of these two metrics, the new paper considers only the page-view statistics of articles about the movies, but extends the sample of movies to include non-American movies as well. Samples of movies in the US, Japan, Australia, the UK, and Germany are studied. The authors concluded: “although the method proposed by Mestyán et al. predicts films’ opening weekend box office revenues in the United States and Australia with reasonable accuracy, its performance drops significantly when applied to various foreign markets. ... we used the model

to predict the opening weekend box office revenues generated by films in British, Japanese, and German theatres, [and] found its accuracy to be far from satisfactory.”

7.0.32 Briefly

“Building academic literacy and research skills by contributing to Wikipedia”

A survey^[9] of research skills of a group of students at an Australian institution showed that purposeful engaging with Wikipedia, including contributing to it, improved their academic skillset.



Map indicating the language areas and provinces of Belgium.

“Google and Bing reintroduce national boundaries more so than Wikipedia does”

In a blog post titled “How does Wikipedia cover the world differently than Google (or Bing)?”,^[10] researcher Han-Teng Liao examines this question by looking at the case of Belgium, which has several language areas. While the two search engines offer a national portal page (google.be / be.bing.com) in different language options, “Wikipedia organizes its users and information less along the lines of national differences and more along the lines of language differences. According to various traffic reports provided by the Wikimedia foundation, users from Belgium contribute to viewing and editing activities mostly in its Dutch, French and English versions.”

7.0.33 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- “Inferring Semantic Facets of a Music Folksonomy with Wikipedia”^[11]

- “Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts”^[12]

- **Pandemic page views in online news media and Wikipedia:** From the English abstract of this German-language paper^[13]: “... a time-series analysis is done comparing the amount of the coverage of eleven online media on the EHEC pandemic in summer 2011 and the amount of page requests for articles in the online encyclopedia Wikipedia relevant to EHEC. Overall, analyses show strong correlations but also temporary discrepancies, appearing because page requests do not only depict the public agenda but also existing uncertainty about an issue.”

- “What Makes a Good Team of Wikipedia Editors? A Preliminary Statistical Analysis”^[14]. From the abstract: “The paper concerns studying the quality of teams of Wikipedia authors with statistical approach. [...] The analysis confirmed that the key issue significantly influencing article’s quality are discussions between team [sic] members. The second part of the paper successfully uses machine learning models to predict good articles based on features of the teams that created them.”

- “A computational linguistic approach towards understanding Wikipedia’s article for deletion (AfD) discussions”^[15]. From the abstract: “In this thesis we aim to solve two main problems: 1) how to help new users effectively participate in the [deletion] discussion; and 2) how to make it efficient for administrators to make decision based on the discussion. To solve the first problem, we obtain a knowledge repository for new users by recognizing imperatives. We propose a method to detect imperatives based on syntactic analysis of the texts. And the result shows a good precision and reasonable recall. To solve the second problem, we propose a decision making support system that provides administrators with an reorganized overview of a discussion.”

7.0.34 References

- [1] Fichman, Pnina (2014). *Global Wikipedia : international and cross-cultural issues in online collaboration*. Lanham: Rowman & Littlefield. ISBN 9780810891012.
- [2] Eom, Young-Ho; Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, Dima L. Shepelyansky (2014-05-28). “Interactions of cultures and top people of Wikipedia from ranking of 24 language editions”. *arXiv: 1405.7183 [physics]*.
- [3] Klein, Max. “Sex Ratios in Wikidata”.
- [4] Mohammad Noor Nawaz and Andrea Bunt (2014) Intel-Wiki: recommending resources to help users contribute to Wikipedia. In *Proceedings of the 22nd International*

Conference on User Modeling, Adaptation, and Personalization (UMAP 2014), 12 pp., forthcoming. PDF

[5]

[6] Liao, Han-Teng (2014-06-17). "What do Chinese-language microblog users do with Baidu Baike and Chinese Wikipedia?".

[7] Solomon, Jacob; Rick Wash (2014-05-16). "Critical mass of what? Exploring community growth in WikiProjects". *Eighth International AAAI Conference on Weblogs and Social Media*. Eighth International AAAI Conference on Weblogs and Social Media.

[8] de Silva, Brian; Ryan Compton (2014-05-22). "Prediction of foreign box office revenues based on wikipedia page activity". *arXiv:1405.5924 [physics]*.

[9] Miller, Julia (2014-06-13). "Building academic literacy and research skills by contributing to Wikipedia: A case study at an Australian university". *Journal of Academic Language and Learning* **8** (2): A72–A86. ISSN 1835-5196.

[10] Liao, Han-Teng (2014-05-13). "How does Wikipedia cover the world differently than Google (or Bing)?".

[11] Sordo, Mohamed; Fabien Gouyon, Luís Sarmiento, Óscar Celma, Xavier Serra (2013). "Inferring semantic facets of a music folksonomy with Wikipedia". *Journal of New Music Research* **42** (4): 346–363. doi:10.1080/09298215.2013.848904. ISSN 0929-8215. Retrieved 2014-06-28.

[12] Joorabchi, Arash; Abdulhussain E. Mahdi (2014-04-01). "Towards linking libraries and Wikipedia: automatic subject indexing of library records with Wikipedia concepts" (PDF). *Journal of Information Science* **40** (2): 211–221. doi:10.1177/0165551513514932. ISSN 0165-5515. Retrieved 2014-06-28.

[13] Holbach, Thomas; Dr Marcus Maurer. "Wissenswertes Nachrichten". *Publizistik*: 1–17. doi:10.1007/s11616-013-0191-z. ISSN 0033-4006.



[14] Bukowski, Leszek; Michał, Jankowski-Lorek, Szymon Jaroszewicz, Marcin Sydow (2014-01-01). "What Makes a Good Team of Wikipedia Editors? A Preliminary Statistical Analysis". In Akiyo Nadamoto, Adam Jatowt, Adam Wierzbicki, Jochen L. Leidner (eds.). *Social Informatics*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 14–28. ISBN 978-3-642-55284-7.

[15] Mao, Wanting. A computational linguistic approach towards understanding Wikipedia's article for deletion (AfD) discussions. Master's thesis The University of Western Ontario, 2014. PDF

Wikimedia Research Newsletter

Vol: 4 • Issue: 6 • June 2014

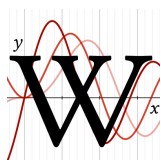
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  **Email**  • [archives] [signpost edition] [contribute] [research index]

Chapter 8

Issue 4(7): July 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 7 • July 2014 [\[contribute\]](#) [\[archives\]](#) 

Shifting values in the paid content debate; cross-language vandalism detection; translations from 53 Wiktionaries

With contributions by: Piotr Konieczny, Maximilian Klein, Heather Ford, and Han-Teng Liao

8.0.35 Understanding shifting values underlying the paid content debate on the English Wikipedia

See related Signpost content: "Extensive network of clandestine paid advocacy exposed", "With paid advocacy in its sights, the Wikimedia Foundation amends their terms of use"

Reviewed by Heather Ford

Kim Osman has performed a fascinating study^[1] on the three 2013 failed proposals to ban paid advocacy editing in the English language Wikipedia. Using a Constructivist Grounded Theory approach, Osman analyzed 573 posts from the three main votes on paid editing conducted in the community in November, 2013. She found that editors who opposed the ban felt that existing policies of neutrality and notability in WP already covered issues raised by paid advocacy editing, and that a fair and accurate encyclopedia article could be achieved by addressing the quality of the edits, not the people contributing the content. She also found that a significant challenge to any future policy is that the community 'is still not clear about what constitutes paid editing'.

Osman uses these results to argue that there has been a transition in the values of the English language Wikipedia editorial community from seeing commercial involvement as direct opposition to Wikipedia's core values (something repeated at the institutional level by the Wikimedia Foundation and Jimmy Wales who see a bright line between paid and unpaid editing) to an acceptance of paid professions and a resignation to their presence.

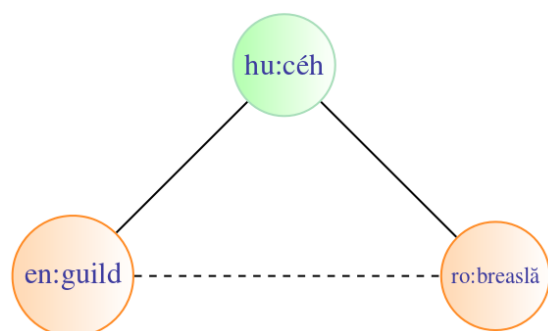
Osman argues that the romantic view of Wikipedia as a system somehow apart from the commercial market that characterized earlier depictions (such as those by Yochai Benkler) has been diluted in recent years and that sustainability in the current environment is linked to a platform's ability to integrate content across multiple places and spaces on the web. Osman also argues that these shifts reflect wider changes in assumptions about commerciality in digital media and that the boundaries between commercial and non-profit in the context of peer production are sometimes fuzzy, overlapping and not clearly defined.

Osman's close analysis of 573 posts is a valuable contribution to the ongoing policy debate about the role of paid editing in Wikipedia and will hopefully be used to inform future debates.

8.0.36 "Pivot-based multilingual dictionary building using Wiktionary"

Reviewed by Maximilian Klein (talk)

To build multilingual dictionaries to and from every language is combinatorially a lot of work. If one uses *triangulation*—if A means B, and B means C, then A means C (see figure)—then a lot of the work can be done by machine. A large closed-source effort did this in 2009^[supp 1], but a new paper by Ács^[2] defends "while our methods are inferior in data size, the dictionaries are available on our website"^[supp 2]. Their approach used the translation tables from 53 Wiktionaries, to make 19 million inferred translations more than the 4 million already occurring in Wiktionary. The researchers steered clear of several classical problems like polysemy, one word having multiple meanings, by using a machine learning classifier. The features used in the classifier were based on the graph-theoretic at-



Straight edges represent translation pairs extracted directly from the Wiktionaries. The pair guild–breaslă was found via triangulating.

tributes of each possible word pair. For instance, if two or more languages can be an intermediate “pivot” language for translation, that turned out to be a good indicator of a valid match. In order to test the precision of these translations, manual spot checking was done and found a precision of 47.9% for newly found word-pairs versus 88.4% for random translations coming out of Wiktionary. As for recall, which tested the coverage of a collection of 3,500 common words, 83.7% of words were accounted for by automatic triangulation in the top 40 languages. That means that right now if we were to try and make a 40-language pocket phrasebook to travel around most of the world just using Wiktionary, about 85% of the time there would be a translation, and it would be between 50–85% correct.

This performance would likely need to increase before any results could be operationalized and contributed back into Wiktionary. However, given the fact that the code used to parse and compare 43 different Wiktionaries was also released on GitHub^[supp 3], that goal is a possibility. It’s yet another testament to the open ecosystem to see a Wikimedia project along with Open Researcher efforts make a resource to rival a closed standard. While Ács’ research isn’t the holy grail of translation between arbitrary languages, it cleverly mixes established theory and open data, and then contributes it back to the community.

8.0.37 “Cross Language Learning from Bots and Users to detect Vandalism on Wikipedia”

Reviewed by Han-Teng Liao (talk)

A new study^[3] by Tran and Christen is the latest example of academic research on vandalism detection which has been developed over the years^[supp 4] in the context of the PAN workshop^[supp 5], where researchers develop both corpus data and tools to uncover plagiarism, authorship, and the misuse of social media/software. This work should be of interests to both researchers and Wikipedians because of (a) the need to detect vandalism and (b) the

interesting question whether such vandalism-fighting data and tools are transferable or portable from one language version to another. Both the vandalism-fighting corpus and tools have both practical and theoretical implications for understanding the cross-lingual transfer in knowledge and bots.

In 2010 and 2011, Wikipedia vandalism detection competitions were included by the PAN as workshops. It started with Martin Potthast’s work on building the free-of-charge PAN Wikipedia vandalism corpus, PAN-WVC-10 for research, which compiled 32452 edits based on 28468 Wikipedia articles, among which 2391 vandalism instances were identified by human coders recruited from Amazon’s Mechanical Turk^[supp 6]. In 2011, a larger crowdsourced corpus of 30,000+ Wikipedia edits is released in three languages: English, German, and Spanish^[supp 7], with 65 features to capture vandalism.

Based on even larger datasets of over 500 million revisions across five languages (en:English, de:German, es:Spanish, fr:French, and ru:Russian), Tran & Christen’s latest work adds to the efforts by applying several supervised machine learning algorithms from the Scikit-learn toolkit^[supp 8], including Decision Tree (DT), Random Forest (RF), Gradient Tree Boosting (GTB), Stochastic Gradient Descent (SGD) and Nearest Neighbour (NN).

What Tran & Christen confirm from their findings is that “distinguishing the vandalism identified by bots and users show statistically significant differences in recognizing vandalism identified by users across languages, but there are no differences in recognizing the vandalism identified by bots” (p.13) This demonstrates human beings can recognize a much wider spectrum of vandalism than bots, but still bots are shown to be trainable to be more sophisticated to capture more and more nonobvious cases of vandalism.

Tran & Christen try to further make the case for the benefits of cross language learning of vandalism. They argue that the detection models are generalizable, based on the positive results of transferring the machine-learned capacity from English to other smaller Wikipedia languages. While they are optimistic, they acknowledge such generalization has at best been proven among some of the languages they studied (these languages are all Roman-alphabet-based languages except for Russian), and the poor performance of the Russian language model. Thus, Tran & Christen rightly point out the need for research on non-English and especially non-European language versions. They also recognize that many word based features are no longer useful for some languages such as Mandarin Chinese, because of tokenization and other language-specific issues.

Tran & Christen call for next research projects to include languages such as Arabic and Mandarin Chinese to complete the United Nations working set of languages. It will be interesting to see how such research projects can be executed and how the greater Wikipedia research and editor

community can help and/or use such research efforts.

8.0.38 Readers' interests differ from editors' preferences

Reviewed by Piotrus.

A conference paper titled “Reader Preferences and Behavior on Wikipedia”^[4] deals with the under-studied population of Wikipedia readers. The paper provides a useful literature review on the few studies about reading preference of that group. The researchers used publicly available page view data, and more interestingly, were able to obtain browsing data (such as time spend by a reader on a given page). Since such data is unfortunately not collected by Wikipedia, the researchers obtained this data through volunteers using a [Yahoo!](#) toolbar. The authors used [Wikipedia:Assessment](#) classes to gauge article's quality.

The paper offers valuable findings, including important insights to the Wikipedia community, namely that “the most read articles do not necessarily correspond to those frequently edited, suggesting some degree of non-alignment between user reading preferences and author editing preference”. This is not a finding that should come as much surprise, considering for example the high percentage of quality military history articles produced by the [WikiProject Military History](#), one of the most active if not *the* most active wikiproject in existence - and of how little importance this topic is to the general population. Statistics on topics popularity and quality of corresponding articles can be seen in [Table 1](#), page 3 of the article. [Figure 1](#) on page 4 is also of interest, presenting a matrix of articles grouped by popularity and length. For example, the authors identify the area of “technology” as the 4th most popular, but the quality of its articles lags behind many other fields, placing it around the 9th place. It would be a worthwhile exercise for the Wikipedia community to identify popular articles that are in need of more attention (through revitalizing tools like [Wikipedia:Popular pages](#), perhaps using code that makes [WikiProject popular pages listing](#) work?) and direct more attention towards what our readers want to read about (rather than what we want to write about). Finally, the authors also identify different reading patterns, and suggest how those can be used to analyze article's popularity in more detail.

Overall, this article seems like a very valuable piece of research for the Wikipedia community and the WMF, and it underscores why we should reconsider collecting more data on our readers' behavior. In order to serve our readers as best as we can, more information on their browsing habits on Wikipedia could help to produce more valuable research like this project.

8.0.39 Wikipedia from the perspective of PR and marketing

Reviewed by Piotrus.

An article^[5] in “[Business Horizons](#)”, written in a very friendly prose (not a common finding among academic works), looks at Wikipedia (as well as some other forms of collaborative, [Web 2.0](#) media) from the business perspective of a public relations/marketing studies. Of particular interest to the Wikipedia community is the authors goal of presenting “the three bases of getting your entry into Wikipedia, as well as a set of guidelines that help manage the potential Wikipedia crisis that might happen one day.” The authors correctly recognize that Wikipedia has policies that must be adhered to by any contributors, though a weakness of the paper is that while it discusses Wikipedia concepts such as [neutrality](#), [notability](#), [verifiability](#), and [conflict of interest](#), it does not link to them. The paper provides a set of practical advice on how to get one's business entry on Wikipedia, or how to improve it. While the paper does not suggest anything outright unethical, it is frank to the point of raising some eyebrows. While nobody can disagree with advice such as “as a rule of thumb, try to remain as objective and neutral as possible” and “when in doubt, check with others on the talk page to determine whether proposed changes are appropriate”, given the lack of consensus among Wikipedia's community on how to deal with for-profit and PR editors, other advice such as “maximize mentions in other Wikipedia entries” (i.e. [gaming WP:RED](#)), “be associated with serious contributors...leverage the reputation of an employee who is already a highly active contributor... [befriend Wikipedians in real life]”, “When correcting negative information is not possible, try counterbalancing it by adding more positive elements about your firm, as long as the facts are interesting and verifiable”, “...you might edit the negative section by replacing numerals (99) with words (ninety-nine), since this is also less likely to be read. Add pictures to draw focus away from the negative content” might be seen as more controversial, falling into the [gaming the system](#) gray area. The “Third, get help from friends and family” section in particular seems to fall foul of [meatpuppetry](#).

In the end, this is an article worth reading in detail by all interested in the PR/COI topics, though for better or worse, the fact that it is closed access will likely reduce its impact significantly. On an ending note, one of the two article's co-authors has a page on Wikipedia at [Andreas Kaplan](#), which was restored by a newbie editor in 2012, two years after it's deletion, has been maintained by throw-away SPAs, and this reviewer cannot help but notice that it still seems to fail [Wikipedia:Notability](#) (academics)...

8.0.40 “No praise without effort: experimental evidence on how rewards affect Wikipedia’s contributor community”

Reviewed by Piotrus.

In 2012, the authors of this paper^[6] have given out over a hundred barnstars to the top 1% most active Wikipedians, and concluded that such awards improve editors productivity. This time they repeated this experiment while broadening their sample size to the top 10% most active editors. After excluding administrators and recently inactive editors, they handed out 300 barnstars “with a generic positive text that expressed community appreciation for their contributions”, divided between the 91st–95th, 96th–99th, and 100th percentiles of the most active editors (this corresponds to an average of 282, 62 and 22 edits per month) and then tracked the activity of those editors, as well as of the corresponding control sample which did not receive any award. The experiment was designed to test the hypothesis that less active contributors will be responsive to rewards, similar to the most highly-active contributors from the prior research.

The authors found, however, that rewarding less productive editors did not stimulate higher subsequent productivity. They note that while the top 1% group responded to an award with an increase in productivity (measured at a rather high 60% increase), less productive subjects did not change their behavior significantly. The researchers also noted that while some of the top 1% editors received an additional award from other Wikipedians, not a single subject from the less active group was a recipient of another award.

The researchers conclude that “this supports the notion that peer production’s incentive structure is broadly meritocratic; we did not observe contributors receiving praise or recognition without having first demonstrated significant and substantial effort.” While this will come as little surprise to the Wikipedia community, their other observation - that outside the top 1% of editors, awards such as barnstars have little meaningful impact - is more interesting.

Further, the authors found that while rewarding the most active editors tends to increase their retention ratio, it may counter-intuitively decrease the retention ratio of the less active editors. The authors propose the following explanation: “Premature recognition of their work may convey a different meaning to these contributors; instead of signaling recognition and status in the eyes of the community, these individuals may perceive being rewarded as a signal that their contributions are sufficient, for the time being, or come to expect being rewarded for their contributions.” They suggest that this could be better understood through future research. For the community in general, it raises an interesting question: how should we recognize less active editors, to make sure that thanking

them will not be taken as “you did enough, now you can leave”?

8.0.41 Briefly

Wikipedia assignments improve students’ research skills

It is refreshing to see a continuing and growing stream of academic works endorsing various aspects of teaching with Wikipedia paradigm. A study^[7] of eleven students “enrolled in a semester-long academic literacy course in a preparatory program for study at an Australian university... showed an educationally statistical improvement in the students’ research skills, while qualitative comments revealed that despite some technical difficulties in using the Wikipedia site, many students valued the opportunity to write for a ‘real’ audience and not just for a lecturer.”

A split in the growing field of Chinese-language Wikipedia research

A blog post^[8] by Han-Teng Liao (???) presents an interesting exploratory overview of a Chinese language research on Wikipedia. The findings suggest that Chinese-language scholars and academic publication outlets are increasingly doing research in the field of Wikipedia studies; however there’s “a divide between mainland Chinese academic sources/search results on one hand, and Hong Kong/Taiwanese ones on the other.” The reason for this seems to be primarily technical, as scholars from different regions seem to publish in different outlets, which in turn are not indexed in the academic search engines preferred by those from other region.

8.0.42 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- “Uneven Openness: Barriers to MENA [Middle East/North Africa] Representation on Wikipedia”^[9] (blog post)
- “Detecting epidemics using Wikipedia article views: A demonstration of feasibility with language as location proxy”^[10]
- “The Reasons of People Continue Editing Wikipedia Content - Task Value Confirmation Perspective”^[11]
- “Circling the Infinite Loop, One Edit at a Time: Seriality in Wikipedia and the Encyclopedic Urge”^[12]

- **“Identifying Duplicate and Contradictory Information in Wikipedia”**^[13]
- **“The impact of elite vs. non-elite contributor groups in online social production communities: The case of Wikipedia”**^[14]
- **“What do we Think an Encyclopaedia is?”**^[15]
From the abstract: “Based on survey and interview research carried out with publishers, librarians and higher education students, [this article] demonstrates that certain physical features and qualities are associated with the encyclopaedia and continue to be valued by them. Having identified these qualities, the article then explores whether they apply to three incidences of electronic encyclopaedias, Britannica Online, The Stanford Encyclopedia of Philosophy and Wikipedia.”
- **“Crowdsourcing Knowledge Interdiscursive Flows from Wikipedia into Scholarly Research”**^[16]. From the abstract: “using a dataset collected from the Scopus research database, which is processed with a combination of bibliometric techniques and qualitative analysis [this article finds] that there has been a significant increase in the use of Wikipedia as a reference within all areas of science and scholarship. Wikipedia is used to a larger extent within areas like Computer Science, Mathematics, Social Sciences and Arts and Humanities, than in Natural Sciences, Medicine and Psychology.”
- **“How Readers Shape the Content of an Encyclopedia: A Case Study Comparing the German Meyers Konversationslexikon (1885-1890) with Wikipedia (2002-2013)”**^[17]

8.0.43 References

- [1] Osman, Kim (2014-06-17). “The Free Encyclopaedia that Anyone can Edit: The Shifting Values of Wikipedia Editors”. *Culture Unbound: Journal of Current Cultural Research* **6**: 593–607. doi:10.3384/cu.2000.1525.146593. ISSN 2000-1525.
- [2] Ács, Judit (May 26–31, 2014). “Pivot-based multilingual dictionary building using Wiktionary” (PDF).
- [3] Tran, Khoi-Nguyen; P. Christen (2014). “Cross Language Learning from Bots and Users to detect Vandalism on Wikipedia”. *IEEE Transactions on Knowledge and Data Engineering*. Early Access Online. doi:10.1109/TKDE.2014.2339844. ISSN 1041-4347.
- [4] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, Andreas Kaltenbrunner: Reader Preferences and Behavior on Wikipedia. HT¹⁴, September 1–4, 2014, Santiago, Chile. <http://www.dcs.gla.ac.uk/~{ }mounia/Papers/wiki.pdf>
- [5] Kaplan, Andreas; Michael Haenlein. “Collaborative projects (social media application): About Wikipedia, the free encyclopedia”. *Business Horizons*. doi:10.1016/j.bushor.2014.05.004. ISSN 0007-6813.
- [6] Restivo, Michael; Arnout van de Rijt. “No praise without effort: experimental evidence on how rewards affect Wikipedia’s contributor community”. *Information, Communication & Society*: 1–12. doi:10.1080/1369118X.2014.888459. ISSN 1369-118X.
- [7] Miller, Julia (2014-06-13). “Building academic literacy and research skills by contributing to Wikipedia: A case study at an Australian university”. *Journal of Academic Language and Learning* **8** (2): A72–A86. ISSN 1835-5196.
- [8] Liao, Han-Teng (2014-06-20). “Chinese-language literature about Wikipedia: a meta-analysis of academic search engine result pages”.
- [9] Graham, Mark; Bernie Hogan (2014-04-29). *Uneven Openness: Barriers to MENA Representation on Wikipedia*. Rochester, NY: Social Science Research Network.
- [10] Generous, Nicholas; Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, Reid Priedhorsky (2014-05-14). “Detecting epidemics using Wikipedia article views: A demonstration of feasibility with language as location proxy”. *arXiv:1405.3612 [physics]*.
- [11] Lai, Cheng-Yu; Heng-Li Yang. “The Reasons of People Continue Editing Wikipedia Content - Task Value Confirmation Perspective”. *Behaviour & Information Technology* (ja): 1–47. doi:10.1080/0144929X.2014.929744. ISSN 0144-929X.
- [12] Salor, E.: Circling the Infinite Loop, One Edit at a Time: Seriality in Wikipedia and the Encyclopedic Urge. In Allen, R. and van den Berg, T. (eds.) *Serialization in Popular Culture*. London: Routledge p.170 ff.
- [13] Weissman, Sarah; Samet Ayhan, Joshua Bradley, Jimmy Lin (2014-06-04). “Identifying Duplicate and Contradictory Information in Wikipedia”. *arXiv:1406.1143 [cs]*.
- [14] Mihai Grigore, Bernadetta Tarigan, Juliana Sutanto and Chris Dellarocas: “The impact of elite vs. non-elite contributor groups in online social production communities: The case of Wikipedia”. SCECR 2014 PDF
- [15] Schopflin, Katharine (2014-06-17). “What do we Think an Encyclopaedia is?”. *Culture Unbound: Journal of Current Cultural Research* **6**: 483–503. doi:10.3384/cu.2000.1525.146483. ISSN 2000-1525.
- [16] Lindgren, Simon (2014-06-17). “Crowdsourcing Knowledge Interdiscursive Flows from Wikipedia into Scholarly Research”. *Culture Unbound: Journal of Current Cultural Research* **6**: 609–627. doi:10.3384/cu.2000.1525.146609. ISSN 2000-1525.
- [17] Spree, Ulrike (2014-06-17). “How Readers Shape the Content of an Encyclopedia: A Case Study Comparing the German Meyers Konversationslexikon (1885-1890) with Wikipedia (2002-2013)”. *Culture Unbound: Journal of Current Cultural Research* **6**: 569–591. doi:10.3384/cu.2000.1525.146569. ISSN 2000-1525.



Supplementary references and notes:

- [1] Mausam and Soderland, Stephen and Etzioni, Oren and Weld, Daniel S. and Skinner, Michael and Bilmes, Jeff (2009). “Compiling a Massive, Multilingual Dictionary via Probabilistic Inference”.
- [2] “Hungarian Front Page”.
- [3] “wiki2dict github”.
- [4] For example, in 2013 only two languages are studied in contrast to the five languages reported in this 2014 journal article.
- [5] <http://pan.webis.de/>
- [6] See
- [7] See
- [8] Scikit-learn is an open source project in Python for machine-learning

Wikimedia Research Newsletter

Vol: 4 • Issue: 7 • July 2014

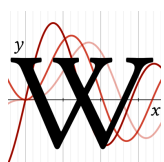
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost

Subscribe:  **Email**  • [archives] [signpost edition]
[contribute] [research index]

Chapter 9

Issue 4(8): August 2014

Wikimedia Research



Vol: 4 • Issue: 8 • August 2014 [contribute] [archives]



A Wikipedia-based Pantheon; new Wikipedia analysis tool suite; how AfC hamstrings newbies

With contributions by: Federico Leva, Piotr Konieczny, Maximilian Klein, and Pine

9.0.44 Wikipedia in all languages used to rank global historical figures of all time

A research group at MIT led by Cesar A. Hidalgo published^[1] a global “Pantheon” (probably the same project already mentioned in our December 2012 issue), where Wikipedia biographies are used to identify and “score” thousands of *global* historical figures of all time, together with a previous compilation of persons having written sources about them. The work was also covered in several news outlets. We won't summarise here all the details, strengths and limits of their method, which can already be found in the well-written document above.

Many if not most of the headaches encountered by the research group lie in the work needed to aggregate said scores by geographical areas. It's easy to get the city of birth of a person from Wikipedia, but it's hard to tell to what ancient or modern country that city corresponds, for any definition of “country”. (Compare our recent review of a related project by a different group of researchers that encountered the same difficulties: “Interactions of cultures and top people of Wikipedia from ranking of



The National Pantheon of the Heroes in Paraguay

24 language editions”). The MIT research group has to manually curate a local database; in an ideal world, they'd just fetch from Wikidata via an API. Aggregation by geographical area, for this and other reasons, seems of lesser interest than the place-agnostic person rank.

The most interesting point is that a person is considered historically relevant when being the subject of an article on 25 or more editions of Wikipedia. This method of assessing an article's importance is often used by editors, but only as an unscientific approximation. It's a useful finding that it proved valuable for research as well, though with acknowledged issues. The study is also one of the rare times researchers bother to investigate Wikipedia in all languages at the same time and we hope there will be follow-ups. For instance, it could be interesting to know which people with an otherwise high “score” were *not* included due to the 25+ languages filter, which could then

be further tweaked based on the findings. As an example of possible distortions, Wikipedia has a dozen subdomains for local languages of Italy, but having an article in 10 italic languages is not an achievement of “global” coverage more than having 1.

The group then proceeded to calculate a “historical cultural production index” for those persons, based on pageviews of the respective biographies (*PV*). This reviewer would rather call it a “historical figures modern popularity index”. While the recentism bias of the Internet (which Wikipedia acknowledges and tries to fight back) for *selection* is acknowledged, most of the recentism in this work is in ranking, because of the usage of pageviews. As *WikiStats* shows, 20% of requests come from a country (the US) with only 5% of the world population, or some 0.3% of the total population in history (assumed as ~108 billion). Therefore there is an error/bias of probably two orders of magnitude in the “score” for “USA” figures; perhaps three, if we add that five years of pageviews are used as sample for the whole current generation. L^* is an interesting attempt to correct the “languages count” for a person (L) in the cases where visits are amassed in single languages/countries; but a similar correction would be needed for *PV* as well.

From the perspective of Wikipedia editors, it’s a pity that Wikipedia is the main source for such a rank, because this means that Wikipedians can’t use it to fill gaps: the distribution of topic coverage across languages is complex and far from perfect; while content translation tools will hopefully help make it more even, prioritisation is needed. It would be wonderful to have a rank of notably missing biographies per language editions of Wikipedia, especially for under-represented groups, which could then be forwarded to the local editors and featured prominently to attract contributions. This is a problem often worked on, from ancient times to recent tools, but we really lack something based on third party sources. We have good tools to identify languages where a given article is missing, but we first need a list (of lists) of persons with *any* identifier, be it authority record or Wikidata entry or English name or anything else that we can then map ourselves.

The customary complaint about inconsistent inclusion criteria can also be found: «being a player in a second division team in Chile is more likely to pass the notoriety criteria required by Wikipedia Editors than being a faculty at MIT», observe the MIT researchers. However, the fact that nobody has bothered to write an article on a subject doesn’t mean that the project as a whole is not interested in having that article; articles about sports people are just easier to write, the project needs and wants more volunteers for everything. Hidalgo replied that he had some examples of deletions in mind; we have not reviewed them, but it’s also possible that the articles were deleted for their state rather than for the subject itself, a difference to which “victims” of deletion often fail to pay attention to.

9.0.45 WikiBrain: Democratizing computation on Wikipedia

– by Maximilianklein

When analyzing any Wikipedia version, getting the underlying data can be a hard engineering task, beyond the difficulty of the research itself. Being developed by researchers from Macalester College and the University of Minnesota, *WikiBrain* aims to “run a single program that downloads, parses, and saves Wikipedia data on commodity hardware.”^[2] Wikipedia dump-downloaders and parsers have long existed, but *WikiBrain* is more ambitious in that it tries to be even friendlier by introducing three main *primitives*: a multilingual concept network, semantic relatedness algorithms, and geospatial data integration. With those elements, the authors are hoping that Wikipedia research will become a mix-and-match affair.



Waldo Tobler’s First Law of Geography – “everything is related to everything else, but near things are more related than distant things” – can be shown true for Wikipedia articles in just a few lines of code with *WikiBrain*.

The first primitive is the multilingual concept network. Since the release of Wikidata, the Universal Concepts that all language versions of Wikipedia represent have mostly come to be defined by the Wikidata item that each language mostly links to. “Mostly” is a key word here, because there are still some edge cases, like the English Wikipedia’s distinguishing between the concepts of “high school” and “secondary school”, while others do not. *WikiBrain* will give you the Wikidata graph of multilingual concepts by default, and the power to tweak this as you wish.

The next primitive is *semantic relatedness* (SR), which is the process of quantifying how close two articles are by their meaning. There have been literally hundreds of SR algorithms proposed over the last two decades. Some rely on Wikipedia's links and categories directly. Others require a text corpus, for which Wikipedia can be used. Most modern SR algorithms can be built one way or another with Wikipedia. WikiBrain supplies the ability to use five state-of-the-art SR algorithms, or their *ensemble* method – a combination of all 5.

Already at this point an example was given of how to mix our primitives. In just a few lines of code, one could easily find which articles in all languages were closest to the English article on “jazz”, and which were also tagged as a film in Wikidata.

The last primitive is a suite of tools that are useful for spatial computation. So extracting location data out of Wikipedia and Wikidata can become a standardized process. Incorporated are some classic solutions to the “geoweb scale problem” – that regardless of an entity's footprint in space, it is represented by a point. That is a problem one shouldn't have to think about, and indeed, WikiBrain will solve it for you *under the covers*.

To demonstrate the power of WikiBrain the authors then provide a case study wherein they replicate previous research that took “thousands of lines of code”, and do it in “just a few” using WikiBrain's high-level syntax. The case study is cherry-picked as is it previous research of one of the listed authors on the paper – of course it's easy to reconstruct one's own previous research in a framework you custom-built. The case study is an empirical testing of *Tobler's first law of geography* using Wikipedia articles. Essentially one compares the SR of articles versus their geographic closeness – and it's verified they are positively linked.

Does the world need an easier, simpler, more off-the-shelf Wikipedia research tool? Yes, of course. Is WikiBrain it? Maybe or maybe not, depending on who you are. The software described in the paper is still version 0.3. There are notes explaining the upcoming features of edit history parsing, article quality ranking, and user data parsing. The project and its examples are written in Java, which is a language choice that targets a specific demographic of researchers, and alienates others. That makes WikiBrain a good tool for Java programmers who do not know how to parse off-line dumps, and have an interest in either multilingual concept alignment, semantic relatedness, and spatial relatedness. For everyone else, they will have to make do with one of the other 20+ alternative parsers and write their own glueing code. That's OK though; frankly the idea to make one research tool to “rule them all” is too audacious and commandeering for the open-source ecosystem. Still that doesn't mean that WikiBrain can't find its userbase and supporters.

9.0.46 Newcomer productivity and pre-publication review

It's time for another interesting paper on newcomer retention^[3] from authors with a proven track record of tackling this issue. This time they focus on the *Articles for Creation* mechanism. The authors conclude that instead of improving the success of newcomers, AfC in fact further decreases their productivity. The authors note that once AfC was fully rolled out around mid-2011, it began to be widely used – the percentage of newcomers using it went up from <5% to ~25%. At the same time, the percentage of newbie articles surviving on Wikipedia went down from ~25% to ~15%. The authors hypothesize that the AfC process is unfriendly to newcomers due to the following issues: 1) it's too slow, and 2) it hides drafts from potential collaborators.

The authors find that the AfC review process is not subject to insurmountable delays; they conclude that “most drafts will be submitted for review quickly and that reviews will happen in a timely manner.”. In fact, two-thirds of reviews take place within a day of submission (a figure that positively surprised this reviewer, though a *current AfC status report* suggests a situation has worsened since: “Severe backlog: 2599 pending submissions”). In either case, the authors find that about a third or so of newcomers using the AfC system fail to understand the fact that they need to finalize the process by submitting their drafts to the review at all – a likely indication that the AfC instructions need revising, and that the AfC regulars may want to implement a system of identifying stalled drafts, which in some cases may be ready for mainspace despite having never been officially “submitted” (due to their newbie creator not knowing about this step or carrying it out properly).

However, the authors do stand by their second hypothesis: they conclude that the AfC articles suffer from not receiving collaborative help that they would get if they were mainspaced. They discuss a specific AfC, for the article *Dwight K. Shellman, Jr/Dwight Shellman*. This article has been tagged as *potentially rescuable*, and has been languishing in that state for years, hidden in the AfC namespace, together with many other *similarly backlogged articles*, all stuck in low-visibility limbo and prevented from receiving proper Wikipedia-style collaboration-driven improvements (or deletion discussions) as an article in the mainspace would receive.

The researchers identify a number of other factors that reduce the functionality of the AfC process. As in many other aspects of Wikipedia, negative feedback dominates. Reviewers are rarely thanked for anything, but are more likely to be criticized for passing an article deemed problematic by another editor; thus leading to the mentality that “rejecting articles is safest” (as newbies are less likely to complain about their article's rejection than experienced editors about passing one). AfC also suffers from the same “one reviewer” problem as GA – the re-

viewer may not always be qualified to carry out the review, yet the newbies have little knowledge how to ask for a second opinion. The authors specifically discuss a case of reviewers not familiar with the specific notability criteria: "[despite being notable] an article about an Emmy-award winning TV show from the 1980's was twice declined at AfC, before finally being published 15 months after the draft was started". Presumably if this article was not submitted to a review it would never be deleted from the mainspace.

The authors are critical of the interface of the AfC process, concluding that it is too unfriendly to newbies, instruction wise: "Newcomers do not understand the review process, including how to submit articles for review and the expected timeframe for reviews" and "Newcomers cannot always find the articles they created. They may recreate drafts, so that the same content is created and reviewed multiple times. This is worsened by having multiple article creation spaces (Main, userspace, Wikipedia talk, and the recently-created Draft namespace".

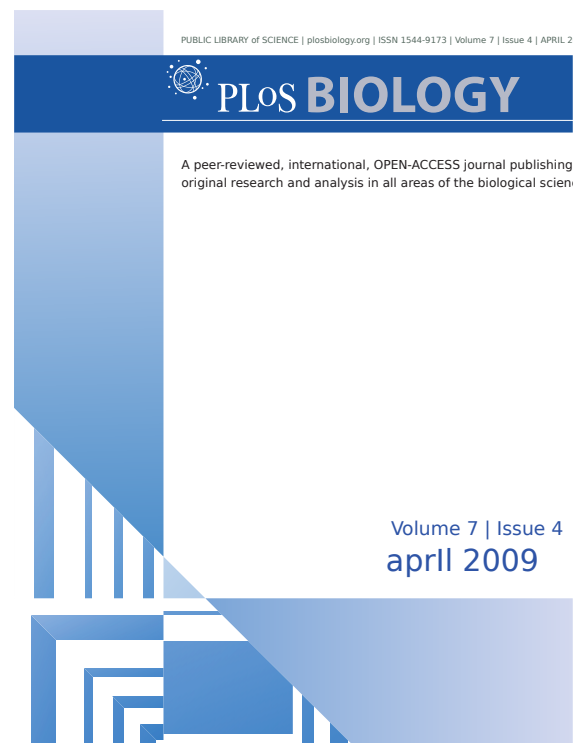
The researchers conclude that AfC works well as a filtering process for the encyclopedia, however "for helping and training newcomers [it] seems inadequate". AfC succeeds in protecting content under the (recently established) speedy deletion criterion G13, in theory allowing newbies to keep fixing it – but many do not take this opportunity. Nor can the community deal with this, and thus the authors call for a creation of "a mechanism for editors to find interesting drafts". That said, this reviewer wants to point out that the G13 backlog, while quite interesting (thousands of articles almost ready for main space ...), is not the only backlog Wikipedia has to deal with – something the writers overlook. The G13 backlog is likely partially a result of imperfect AfC design that could be improved, but all such backlogs are also an artifact of the lack of active editors affecting Wikipedia projects on many levels.

In either case, AfC regulars should carefully examine the authors suggestions. This reviewer finds the following ideas in particular worth pursuing. 1) Determine which drafts need collaboration and make them more visible to potential editors. Here the authors suggest use of a recent academic model that should help automatically identify valuable articles, and then feeding those articles to **SuggestBot**. 2) Support newcomers' first contributions – almost a dead horse at this point, but we know we are not doing enough to be friendly to newcomers. In particular, the authors note that we need to create better mechanisms for newcomers to get help on their draft, and to improve the article creation advice – especially the **Article Wizard**. (As a teacher who has introduced hundreds of newcomers to Wikipedia, this reviewer can attest that the current outreach to newbies on those levels is grossly inadequate.)

A final comment to the community in general: was AfC intended to help newcomers, or was it intended from the start to reduce the strain on new page patrollers by sand-

boxing the drafts in the first place? One of the roles of AfC is to prevent problematic articles from appearing in the mainspace, and it does seem that in this role it is succeeding quite well. English Wikipedia community has rejected the **flagged revisions**-like tool, but allowed implementation of it on a voluntary basis for newcomers, who in turn may not often realize that by choosing the AfC process, friendly on the surface, they are in fact slow-tracking themselves, and inviting extraordinary scrutiny. This leads to a larger question that is worth considering: we, the Wikipedia community of active editors, have declined to have our edits classified as second-tier and hidden from the public until they are reviewed, but we are fine pushing this on to the newbies. To what degree is this contributing to the general trend of Wikipedia being less and less friendly to newcomers? Is the resulting quality control worth turning away potential newbies? Would we be here if years ago our first experience with Wikipedia was through AfC?

9.0.47 Briefly



PLOS Biology is an open-access peer-reviewed scientific journal covering all aspects of biology. Publication began on October 13, 2003.

15% of PLOS Biology articles are cited on Wikipedia

A conference paper titled "An analysis of Wikipedia references across PLOS publications"^[4] asked the following research questions: "1) To what extent are scholarly articles referenced in Wikipedia, and what content is par-

ticularly likely to be mentioned?" and "2) How do these Wikipedia references correlate with other article-level metrics such as downloads, social media mentions, and citations?". To answer this, the authors analyzed which PLOS articles are referenced on Wikipedia. They found that as of March 2014, about 4% of PLOS articles were mentioned on Wikipedia, which they conclude is "similar to mentions in science blogs or the post-publication peer review service, F1000Prime". About half of articles mentioned on Wikipedia are also mentioned on Facebook, suggesting that being cited on Wikipedia is related to being picked up by other social media. Most of Wikipedia cites come from PLOS Genetics, PLOS Biology and other biology/medicine related PLOS outlets, with PLOS One accounting for only 3% total, though there are indications this is changing over time. 15% of all articles from PLOS Biology have been cited on Wikipedia, the highest ratio among the studied journals. Unfortunately, this is very much a descriptive paper, and the authors stop short of trying to explain or predict anything. The authors also observe that "By far the most referenced PLOS article is a study on the evolution of deep-sea gastropods (Welch, 2010) with 1249 references, including 541 in the Vietnamese Wikipedia."

"Big data and small: collaborations between ethnographers and data scientists"

Ethnography is often seen as the least quantitative branch of social science, and this^[5] essay-like article's style is a good illustration. This is, essentially, a self-reflective story of a Wikipedia research project. The author, an ethnographer, recounts her collaboration with two big data scholars in a project dealing with a large Wikipedia dataset. The results of their collaboration are presented here and have been briefly covered by our Newsletter in Issue 8/13. This article can be seen as an interesting companion to the prior, Wikipedia-focused piece, explaining how it was created, though it fails to answer questions of interest to the community, such as "why did the authors choose Wikipedia as their research ground" or about their experiences (if any) editing Wikipedia.

"Emotions under discussion: gender, status and communication in online collaboration"

Researchers investigated^[6] "how emotion and dialogue differ depending on the status, gender, and the communication network of the ~12,000 editors who have written at least 100 comments on the English Wikipedia's article talk pages." Researchers found that male administrators tend to use an impersonal and neutral tone. Non-administrator females used more relational forms of communication. Researchers also found that "editors tend to interact with other editors having similar emotional styles (e.g., editors expressing more anger connect more with one another)." Authors of this paper will present their re-

search at the September Wikimedia Research and Data showcase.



9.0.48 References

- [1] <http://pantheon.media.mit.edu/methods>
- [2] Sen, Shilad; Jia-Jun Li, Toby; WikiBrain Team; Hecht, Brent. "WikiBrain: Democratizing computation on Wikipedia" (PDF). *OpenSym '14* 0 (0): 1–19. doi:10.1145/2641580.2641615.
- [3] Jodi Schneider, Bluma S. Gelley Aaron Halfaker: Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review <http://jodischneider.com/pubs/opensym2014.pdf> OpenSym '14 , August 27–29, 2014, Berlin
- [4] Fenner, Martin; Jennifer Lin (June 6, 2014), "An analysis of Wikipedia references across PLOS publications", *altmetrics14 workshop at WebSci*, doi:10.6084/m9.figshare.1048991
- [5] Ford, Heather (1 July 2014). "Big data and small: collaborations between ethnographers and data scientists". *Big Data & Society* 1 (2): 2053951714544337. doi:10.1177/2053951714544337. ISSN 2053-9517.
- [6] Laniado, David; Carlos Castillo; Mayo Fuster Morell; Andreas Kaltenbrunner (2014-08-20). "Emotions under Discussion: Gender, Status and Communication in Online Collaboration". *PLoS ONE* 9 (8): e104880. doi:10.1371/journal.pone.0104880.

Wikimedia Research Newsletter

Vol: 4 • Issue: 8 • August 2014

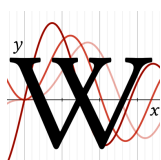
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  [Email](#)  • [\[archives\]](#) [\[signpost edition\]](#) [\[contribute\]](#) [\[research index\]](#)

Chapter 10

Issue 4(9): September 2014

Wikimedia Research



Vol: 4 • Issue: 9 • September 2014 [contribute]
[archives] 

99.25% of Wikipedia birthdates accurate; focused Wikipedians live longer; merging WordNet, Wikipedia and Wiktionary

With contributions by: Scott Hale, Piotr Konieczny, Maximilian Klein, Andrew Krizhanovsky, Tilman Bayer and Pine

10.049 “Reliability of user-generated data: the case of biographical data in Wikipedia”

Review by User:Maximilianklein

0.75% of Wikipedia birthdates are inaccurate, reported Robert Viseur at WikiSym 2014.^[1] Those inaccuracies are “low, although higher than the 0.21% observed for the baseline reference sources”. Given that biographies represent 15% of English Wikipedia,^[supp 1] the third largest category after “arts” and “culture”, their accuracy is important. The method used was to find biographies that were both in Wikipedia and 9 reference databases, which are sadly not named due to the wishes of an “anonymous sponsor” of the paper (Red flag or Belgian bureaucracy?). Of 938 such articles found, those whose birthdates did not match in all 10 databases – 14.4% – were manually investigated. Some errors were due to coincidental names, thus proving the point for authority control in collecting data. One capping anecdote is that most of the mistakes in Wikipedia’s 0.75% were corrected in the intervening



“Third Volume of a 1727 edition of Plutarch’s Lives of the Noble Greeks and Romans printed by Jacob Tonson”; caption quoted from the Wikipedia article Biography

time between data collection and manual investigation. However, one may need to account for the sample bias that these were the biographies which existed in 10 separated databases – well known personalities. Therefore the predictive power of the study remains limited, but at least we know that some objective data on Wikipedia has the same order of magnitude error rate as other “reliable sources”.

10.050 Focused Wikipedians stay active longer

A new preprint^[2] by three Dublin-based computer scientists contributes to the debate around editor retention. The authors use techniques such as the **topic modeling** and **non-negative matrix factorization**, to categorize Wikipedians into several profiles (“e.g. content experts, social networkers”). Those profiles, or user roles, are based on namespaces that editors are most active in. The authors analyzed the behavior of about half a million Wikipedia editors. The authors find that short-term editors seem to lack interest in any one particular aspect of Wikipedia, editing various namespaces briefly before leaving the project. Long-term editors are more likely



Group photo of Wikimedians at Wikimania 2012

to focus on one or two namespaces (usually mainspace, plus article talk or user talk pages), and only after some time diversify to different namespaces; in other words, the namespace distribution of edits over time “predicts an editor’s departure from the community”. The authors note that “we show that understanding patterns of change in user behavior can be of practical importance for community management and maintenance”.

Unfortunately, the paper is heavy in jargon and statistical models, and provides little practical data (or at least, that data is not presented well). For example, the categorization of editors into seven groups is very interesting, but no descriptive data is presented that would allow us to compare the number of editors in each group. Further, the paper promises to use those profiles to predict editor lifecycles, but such models don't seem to be present in the paper. In the end, *this reviewer* finds this paper to be an interesting idea that hopefully will develop into some research with meaningful findings – for now, however, it seems more of a theoretical analysis with no practical applications.

10.0.51 “WordNet-Wikipedia-Wiktionary: construction of a three-way alignment”

Reviewed by Andrew Krizhanovsky

The authors of this paper,^[3] presented at the International Conference on Language Resources and Evaluation (LREC 2014), integrated two previously constructed alignments for WordNet-Wikipedia and WordNet-Wiktionary into a three-way alignment WordNet-Wikipedia-Wiktionary. This integration results in lower accuracy, but greater coverage in comparison with two-way alignment.

Wiktionary does not provide a convenient and consistent means of directly addressing individual lexical items or their associated senses. Third-party tools such as the JWKTLL (Java-based Wiktionary Library) API can overcome this problem.



Wiktionary

The free dictionary

A Wiktionary logo

Since the WordNet–Wikipedia alignment is for nouns only, the resulting synonym sets in the conjoint three-way alignment consist entirely of nouns. However, the full three-way alignment contains all parts of speech (adjectives, nouns, adverbs, verbs, etc.).

Larger synonym sets in the source data (WordNet and Wiktionary) results in more incorrect mapping in the outcome alignment (this is strange from the average person’s point of view and shows that the alignment algorithm is not perfect yet).

Informal examination shows that conjoint alignment is correct in general, but existing errors in the source alignments were magnified (snowball effect).

10.0.52 Briefly

Measures of edit quality

A work-in-progress paper^[4] reviews measures of edit quality on Wikipedia and reports the results of a pilot project to evaluate the “Persistent Word Revisions” (PWR)^[supp 2] metric of edit quality with the ratings of Amazon’s Mechanical Turk users. PWR measures how much of an edit is preserved through subsequent revisions to the article. The paper only evaluates “a small pool of 63 total [Mechanical Turk] ratings of 10 [article] revisions” and therefore has no significant results. Nonetheless, the future validation on a much larger set of edits as promised in the paper should be useful to future researchers. It will also be useful to know how the distribution of PWR scores compare with other measures of article quality such as the quality assessments given by

WikiProjects, nominations for Good Article or Featured Article status. A comparison with Adler et al.'s WikiTrust scores could also be valuable.

“A Wiki Framework for the Sweble Engine”

This master thesis^[5] builds on previous work of professor Dirk Riehle's research group at the University of Erlangen-Nuremberg which had constructed a formal parser for MediaWiki wikitext, adding a web application that allows editing wikis based on this parser.

How quickly are drug articles updated after FDA warnings?

A short article^[6] in the *New England Journal of Medicine* examined how quickly safety warnings by the US Food and Drug Administration (FDA) for 22 prescription drugs were incorporated into the corresponding Wikipedia articles. The authors “found that 41% of Wikipedia pages pertaining to the drugs with new safety warnings were updated within 2 weeks ... The Wikipedia pages for drugs that were intended for treatment of highly prevalent diseases (affecting more than 1 million people in the United States) were more likely to be updated quickly (58% were updated within 2 weeks) than were those for drugs designed to treat less-prevalent conditions (20% were updated within 2 weeks ...)”. See also the discussion at WikiProject Medicine: 1 2

“Spiral of silence” in German Wikipedia's image filter discussions

A paper titled “The Dispute over Filtering 'indecent' Images in Wikipedia”^[7] examines disputes in 2010 and 2011 about controversial content on Wikipedia, and about the Wikimedia Foundation's proposal for an opt-in image filter which would have allowed users to hide sexual or violent media for themselves (see the *Signpost* summary by this reviewer). The author finds that several of German sociologist Jürgen Habermas' criteria for public discourse apply to the lengthy discussions on the German Wikipedia about this topic (highlighting one talk page with 120 major threads that fill 175 pages in a PDF). “However, [Habermas'] criteria of rationality and objectivity seem to be less applicable. Compared to other areas of dispute in Wikipedia, the German discussions were civilized – but emotional.” The paper invokes the “spiral of silence” theory of public opinion to explain the German Wikipedia's huge opposition to the Wikimedia Foundation's plans: “the climate of opinion in the on-line discussions put supporters of the image filter under heavy pressure to conform or to be silent”. Finally, the paper reports on the results of a small web-based experiment where 163 participants were randomly shown one of three versions of the article de:Furunkel (boil): Ei-

ther without images, or with a “neutral image”, or “with a somewhat disgusting image of an infected boil.” The author states that “The most interesting results for the Wikipedia community is that the disgusting image enhances the perceived quality of the article: It is perceived to be more fascinating ($p=.023$) and more worth reading ($p=.032$) than an article without any image.”

10.0.53 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- **“Evolution and revolution of organizational configurations on wikipedia: A longitudinal network analysis”**^[8] From the abstract: “A new step-wise regression model-selection approach was used to detect significant shifts in the trends of in-bound degree centralization, outbound degree centralization, betweenness centralization, assortativity, and social entropy [in the coauthorship network of editors and articles]. ... Finally, the moments of revolutionary change were compared with prominent media stories, news items referencing Wikipedia, and important policy changes and events on Wikipedia...”
- **“Field experiments of success-breeds-success dynamics”**^[9] (coverage of earlier related papers by two of the authors: “Recognition may sustain user participation”, “No praise without effort: experimental evidence on how rewards affect Wikipedia's contributor community”)
- **“How collective intelligence emerges: knowledge creation process in Wikipedia from microscopic viewpoint”**^[10]
- **“How accurate are Wikipedia articles in health, nutrition, and medicine?”**^[11]
- **“Community and the dynamics of spatially distributed knowledge production. The case of Wikipedia”**^[12]
- **“Group minds and the case of Wikipedia”**^[13] (see also coverage of an earlier paper by the author: “Wikipedia editing patterns are consistent with a non-finite state model of computation”)
- **“‘The sum of all human knowledge': A systematic review of scholarly research on the content of Wikipedia”**^[14] (see also mailing list announcement and our coverage of a related paper by the same authors: “Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership”)

10.0.54 References

- [1] VISEUR, Robert (2014). “Reliability of User-Generated Data: the Case of Biographical Data in Wikipedia” (PDF). *WikiSym 2014*. Retrieved 24 September 2014.
- [2] Qin, Xiangju; Derek Greene; Pádraig Cunningham (29 July 2014). “A latent space analysis of editor lifecycles in Wikipedia”. *Proc. of 5th International Workshop on Mining Ubiquitous and Social Environments (MUSE) at ECML/PKDD 2014*. arXiv:1407.7736.
- [3] Miller, Tristan; Iryna Gurevych (May 2014). “WordNet-Wikipedia-Wiktionary: construction of a three-way alignment” (PDF). *Proceedings of the 9th International Conference on Language Resources and Evaluations*. data
- [4] Biancani, Susan (2014). “Measuring the Quality of Edits to Wikipedia” (PDF). *WikiSym 2014*. Retrieved 24 September 2014.
- [5] Liping Wang: A Wiki Framework for the Sweble Engine. Master thesis, Friedrich-Alexander University Erlangen-Nürnberg 2014 PDF
- [6] Hwang, Thomas J.; Florence T. Bourgeois; John D. Seeger (2014). “Drug Safety in the Digital Age”. *New England Journal of Medicine* **370** (26): 2460–2462. doi:10.1056/NEJMp1401767. ISSN 0028-4793. PMID 24963564.
- [7] Thomas Roessing: The Dispute over Filtering “indecent” Images in Wikipedia. Masaryk University Journal of Law and Technology Issue: 2/2013
- [8] Britt, Brian C. (January 2014). “Evolution and revolution of organizational configurations on wikipedia: A longitudinal network analysis”. Purdue University.
- [9] Rijt, Arnout van de; Soong Moon Kang; Michael Restivo; Akshay Patil (28 April 2014). “Field experiments of success-breeds-success dynamics”. *Proceedings of the National Academy of Sciences*: 201316836. doi:10.1073/pnas.1316836111. ISSN 0027-8424. PMID 24778230.
- [10] Lee, Kyungho (2014). “How collective intelligence emerges: knowledge creation process in Wikipedia from microscopic viewpoint”. *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. AVI '14. New York, NY, USA: ACM. pp. 373–374. doi:10.1145/2598153.2600040. ISBN 978-1-4503-2775-6.
- [11] Temple, Norman J.; Joy Fraser (2014). “How accurate are Wikipedia articles in health, nutrition, and medicine? / Les articles de Wikipédia dans les domaines de la santé, de la nutrition et de la médecine sont-ils exacts ?”. *Canadian Journal of Information and Library Science* **38** (1): 37–52. ISSN 1920-7239.
- [12] Joanne Robert: Community and the dynamics of spatially distributed knowledge production. The case of Wikipedia in: *The social dynamics of innovation networks*. edited by Roel Rutten, Paul Bennenworth, Dessy Irawati, Frans Boekema p.179ff
- [13] DeDeo, Simon (8 July 2014). “Group minds and the case of Wikipedia”. arXiv:1407.2210.
- [14] Mesgari, Mostafa and Okoli, Chitu and Mehdi, Mohamad and Nielsen, Finn Årup and Lanamäki, Arto (2014) “The sum of all human knowledge”: A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*. ISSN 2330-1635 (In Press) PDF



Supplementary references and notes:

- [1] Kittur. “Whats in Wikipedia?” (PDF).
- [2] Halfaker, A., Kittur, A., Kraut, R., & Riedl, J. (2009). A Jury. “A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia”. *WikiSym '09*. Retrieved 24 September 2014.

Wikimedia Research Newsletter

Vol: 4 • Issue: 9 • September 2014

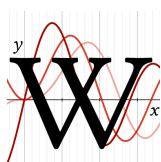
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost

Subscribe:  [Email](#)  • [archives] [signpost edition] [contribute] [research index]

Chapter 11

Issue 4(10): October 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 10 • October 2014 [contribute] [archives]



Informed consent and privacy; newsmaking on Wikipedia; Wikipedia and organizational theories

With contributions by: Maximilian Klein, Piotr Konieczny, Kim Osman, Pine and Tilman Bayer

11.055 **Tl;dr: Users, informed consent and privacy policies online**

Reviewed by Kim Osman

In new research^[1] conducted in light of proposed changes to data protection legislation in the European Union (EU), authors Bart Custers, Simone van der Hof, and Bart Schermer conducted a comparative analysis of social media and user-generated content websites' privacy policies along with a user survey (N=8,621 in 26 countries) and interviews in 13 different EU countries on awareness, values, and attitudes toward privacy online. The authors state consent regarding personal data use is an important concept and observe, "There is mounting evidence that data subjects do not fully contemplate the consequences and risks of personal data processing."

Custers, van der Hof and Schermer developed a set of criteria for giving informed consent about the use of personal data including: "Is it clear who is processing the data and who is accountable?" and "Is the information provided understandable?" When existing privacy policies were applied to these criteria, Wikipedia was the worst performing of the sites analyzed and recommends

that it makes clear how minors are dealt with and to provide additional clarity around security measures. It also notes that IP addresses may be traced, therefore making "anonymous" Wikipedia users identifiable.

The study did acknowledge issues around self-presentation and identity in different online contexts and the actual need for a site like Wikipedia to have an extensive privacy policy as users afford criteria regarding privacy different value in these different contexts. The authors do note however, "Wikipedia does collect opinions that may be attributable to individuals and that may be considered privacy sensitive."

This paper is a well-researched summary of the privacy policies of online sites (including major international platforms like Facebook, Twitter and YouTube), and although from a European perspective (where data collection practices are arguably more stringent than in other places in the world), it raises important questions about how Wikipedia approaches its privacy policy in terms of informed user consent, and would be useful reading for anyone with an interest in how online practices are shaping approaches to user privacy.

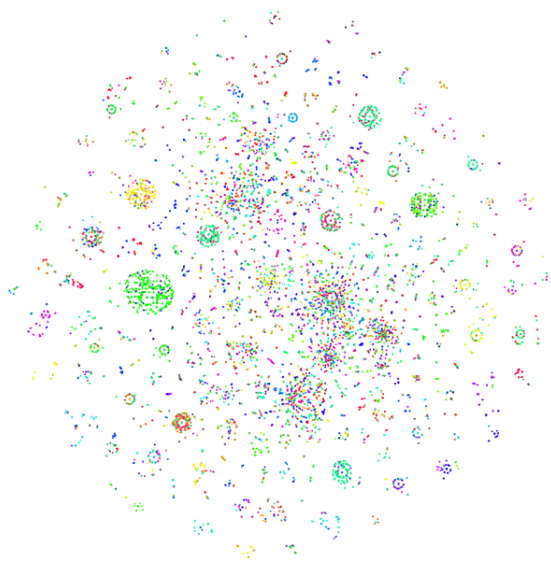
For researchers requiring more information about ethics in online research visit the Association of Internet Researchers' wiki.

11.056 **Briefly**

Holocaust articles compared across languages

We tell ourselves that Wikipedia works well for the most part, but that finding consensus might break down on controversial articles. Of all article topics, perhaps none is potentially more fraught than the Holocaust, and that is precisely what Rudolf Den Hartogh has tackled in his Master's thesis "The future of the Past: A case study on the representation of the Holocaust on Wikipedia".^[2] It is an in-depth *compare and contrast* analysis of the Holocaust topic in the English, German, and Dutch. Several curious facts come out of this. For instance the average vandalism rate on these articles is 4%, compared with 7% globally - as these articles have been locked at some point, although the Dutch version is no longer protected.

Other analyses show edit activity over time, since the articles' inception. The German version saw the height of its shaping 2 years after it was started in 2004, whereas the English and Dutch articles saw their main spurts 5 and 3 years later respectively. Moreover the author finds "that there does not exist one representation of the Holocaust, but each language version has its own unique account of events and phenomena." Finally they "found that none of the Holocaust entries under study is rated 'good quality'," so we still have not definitively addressed the hardest parts of our encyclopedia.



Semantic role label features for all records, colours are based on event tag in the Lensing Wikipedia dataset.

Lensing Wikipedia

A project^[3] with this title aims to extract date, location, event and role semantic data from historical English Wikipedia articles. Of course making grand sense of that automatic extraction work requires visualization. Such visualization is difficult on *high-dimensional data* consisting of e.g. a date, location, multiple events and roles - all at the same time. A short proof of concept "Visualizing Wikipedia using t-SNE" by Jasneet Singh Sabharwal^[4] has done just this using a Barnes-Hut simulation variation of the T-distributed stochastic neighbor embedding algorithm. This image shows the closeness of the semantic roles of features found in Wikipedia article text, with colors indicating similar *events* that articles are describing.

"Infoboxes and cleanup tags: Artifacts of Wikipedia newsmaking"

An article^[5] in *Journalism: Theory, Practice and Criticism* looks at use and abuse of cleanup tags and infobox elements as conceptual and symbolic tools. Based on

ethnographic observations and several interviews, the author provides a lengthy description of the formative first three or so weeks in the 2011 Egyptian Revolution article. It is a valuable study of how articles are developed, and the collaboration and conflicts that are common in high-activity articles. The author provides a valuable observation that "Classification work... is intensely political" and "the editing of Wikipedia articles involves continuous linking and classifying." The choice of words, categories, article titles, but also specific tags or infoboxes (though a particular example discussed - whether to use `Template:Infobox uprising` or not - seems to concern a template that does not, in fact, exist) can be quite controversial. The author also puts forth an interesting argument that removal of cleanup tags may give false impressions of stability in articles that are not yet stable; and that infoboxes carry significant, perhaps undue weight, compared to other elements of the article.

Wikipedia's identity "based on freedom"

This paper^[6] looks at Wikipedia through a number of organizational theory lenses, in particular theories of organizational identity. Of particular interest to Wikipedians is one of the aspects analyzed by the editors - identify of the project. The authors state that "the organizational identity at Wikipedia is based on freedom". Next, they discuss the utopian ideals of freedom (such as "anyone can edit"), as contrasted with the freedom-reducing tendencies of censorship, administrative control, and bureaucratization. The authors argue that the common solution to criticism of Wikipedia, within the community, is concealment and marginalization of said criticism. The authors point to the practical defanging of the `Wikipedia:Ignore all rules` policy, which has went through a number of meaning shifts, in which it was re-defined to be virtually toothless, even though the name remained the same. Another way that freedom is limited is through end-justifies-the-mean utopian vision of "free access [to Wikipedia] for everyone", replacing the older "anyone can edit" "freedom of editing meaning. Unfortunately, the author's discussion of "the subjugation of contesting voices" is very short on details and specifics; the authors allude to administrator power abuse, but fail to provide any specific discussion of how it occurs; an example they used of "deleted content" can be interpreted as nothing more sinister than admin ability to delete content that does not meet Wikipedia's site policies, including uncontroversial content such as spam.

"Copyright or Copyleft? Wikipedia as a Turning Point for Authorship"

This paper^[7] touches upon a very interesting yet understudied area: what Wikipedia's existence means for copyright law. As the authors note, Wikipedia "appears to challenge some of the notions at the heart of copyright

law.”

Critique of Wikipedia’s dispute resolution procedures

This paper^[8] claims to present an ethnographic analysis of and a strong critique of Wikipedia’s dispute resolution procedures, and states upfront its goal as “to tease out systemic discrimination or injustice”. The strongly worded abstract is attention-drawing, promising that “A number of flaws will be identified including the ability for vocal minorities to dominate the Wikipedia community consensus”. Unfortunately, while the paper provides a very detailed description of Wikipedia’s dispute resolution scene, it doesn’t seem to present any new data; its critique of “vocal minorities”, for example, is composed of few sentences, and the entire argument is based on, and essentially a repetition of a similar passage in Reagle’s *Good Faith Collaboration* book. While the paper is well written and presents a number of valid arguments, it does not seem to contribute anything new to our understanding of Wikipedia, being in essence a literature review focused on the topic of dispute resolution on Wikipedia. Which this reviewer finds disappointing, considering that the almost tabloid-style abstract and the introductory section promise ethnographic research, which - like anything else going beyond synthesis of existing, published research - is sadly very much absent from the paper.

11.0.57 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- **“Insights from the Wikipedia Contest (IEEE Contest for Data Mining 2011)”^[9]** (earlier coverage: "Predicting editor survival: The winners of the Wikipedia Participation Challenge")
- **“A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection”^[10]** (constructs a dispute corpus from Wikipedia talk pages)
- **“Extracting Imperatives from Wikipedia Article for Deletion Discussions”^[11]** (without conclusions or published dataset, apparently)
- **“Use of Wikipedia by Legal Scholars: Implications for Information Literacy”^[12]**
- **“Guiding Students in Collaborative Writing of Wikipedia Articles – How to Get Beyond the Black Box Practice in Information Literacy Instruction”^[13]** (received the EdMedia Outstanding Paper Award)
- **“Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project”^[14]** (project

home page, allowing the live creation of a taxonomy graph for an arbitrary Wikipedia article: <http://wibitaxonomy.org>)

- **“Analysis of the accuracy and readability of herbal supplement information on Wikipedia”^[15]**
- **“Maturity Assessment of Wikipedia Medical Articles”^[16]**
- **“Computer-supported collaborative accounts of major depression: Digital rhetoric on Quora and Wikipedia”^[17]**

11.0.58 References



- [1] Custers, Bart; Simone van der Hof, Bart Schermer (2014-09-01). “Privacy Expectations of Social Media Users: The Role of Informed Consent in Privacy Policies”. *Policy & Internet* **6** (3): 268–295. doi:10.1002/1944-2866.POI366. ISSN 1944-2866.
- [2] Den Hartogh, Rudolf (2014). *The future of the Past: A case study on the representation of the Holocaust on Wikipedia* (Masters). Erasmus University Rotterdam.
- [3] “Lensing Wikipedia”. Simon Fraser University Natural Language Laboratory.
- [4] Jasneet Singh Sabharwal: Visualizing Wikipedia using t-SNE
- [5] Ford, Heather (2014-08-31). “Infoboxes and cleanup tags: Artifacts of Wikipedia newsmaking”. *Journalism*: 1464884914545739. doi:10.1177/1464884914545739. ISSN 1741-3001 1464-8849, 1741-3001 Check lissn= value (help).
- [6] Kozica, Arjan M. F.; Christian Gebhardt, Gordon Müller-Seitz, Stephan Kaiser (2014-10-13). “Organizational Identity and Paradox An Analysis of the 'Stable State of Instability' of Wikipedia’s Identity”. *Journal of Management Inquiry*: 1056492614553275. doi:10.1177/1056492614553275. ISSN 1552-6542 1056-4926, 1552-6542 Check lissn= value (help).
- [7] Simone, Daniela (2013-07-01). *Copyright or Copyleft? Wikipedia as a Turning Point for Authorship*. Rochester, NY: Social Science Research Network.
- [8] Ross, Sara (2014-03-01). *Your Day in 'Wiki-Court': ADR, Fairness, and Justice in Wikipedia’s Global Community*. Rochester, NY: Social Science Research Network.
- [9] Desai, Kalpit V.; Roopesh Ranjan (2014-01-07). “Insights from the Wikipedia Contest (IEEE Contest for Data Mining 2011)”. *arXiv:1405.7393 [physics, stat]*.
- [10] Lu Wang, Claire Cardie: A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 693–699, Baltimore, Maryland, USA, June 23–25 2014

- [11] Fiona Mao, Robert E. Mercer, Lu Xiao: Extracting Imperatives from Wikipedia Article for Deletion Discussions Proceedings of the First Workshop on Argumentation Mining, pages 106–107, Baltimore, Maryland USA, June 26, 2014.
- [12] Darryl Maher: Use of Wikipedia by Legal Scholars: Implications for Information Literacy. Master’s thesis, School of Information Management, Victoria University of Wellington, submitted June 2014
- [13] Sormunen, E. & Alamettälä, T. (2014). Guiding Students in Collaborative Writing of Wikipedia Articles – How to Get Beyond the Black Box Practice in Information Literacy Instruction. In: EdMedia 2014 – World Conference on Educational Media and Technology. Tampere, Finland: June 23-26, 2014
- [14] Flati, Tiziano; Daniele Vannella, Tommaso Pasini, Roberto Navigli (2014). “Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project”. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 945–955.
- [15] Phillips, Jennifer; Connie Lam, Lisa Palmisano (2014-07-01). “Analysis of the accuracy and readability of herbal supplement information on Wikipedia”. *Journal of the American Pharmacists Association* **54** (4): 406–414. doi:10.1331/JAPhA.2014.13181. ISSN 1544-3191.
- [16] Conti, Riccardo; Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, Marinella Petrocchi (2014). “Maturity Assessment of Wikipedia Medical Articles”. *Proceedings of the 2014 IEEE 27th International Symposium on Computer-Based Medical Systems. CBMS '14*. Washington, DC, USA: IEEE Computer Society. pp. 281–286. doi:10.1109/CBMS.2014.69. ISBN 978-1-4799-4435-4.
- [17] Rughinis, Cosima; Bogdana Huma, Stefania Matei, Razvan Rughinis (June 2014). *Computer-supported collaborative accounts of major depression: Digital rhetoric on Quora and Wikipedia*. 2014 9th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6. doi:10.1109/CISTI.2014.6876968.

Wikimedia Research Newsletter

Vol: 4 • Issue: 10 • October 2014

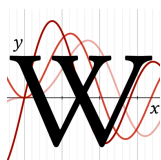
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  [Email](#)  • [\[archives\]](#) [\[signpost edition\]](#)
[\[contribute\]](#) [\[research index\]](#)

Chapter 12

Issue 4(11): November 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 11 • November 2014 [\[contribute\]](#)
[\[archives\]](#) 

Gender gap and skills gap; academic citations on the rise; European food cultures

With contributions by: Piotr Konieczny, Maximilian Klein and Tilman Bayer.

12.0.59 “Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia”

This article^[1] contributes to the discussion on gender inequalities on Wikipedia. The authors take a novel approach of looking for answers outside the Wikipedia community, thus also tying their research into the analysis of new editors recruitment, motivations, and barriers to contribute. The authors focus their analysis on the role of Internet experiences and skills, and their lack among certain groups. The authors study whether the level of one’s skills in digital literacy is related to their chance of becoming a Wikipedia editor, by surveying 547 young adults (aged 21–22) – students at a (presumably American) university, the most used convenience sample in academia. The survey was carried out in 2009, with a follow-up wave in 2012. The students were asked about their socioeconomic and demographic background, as well as about their level of digital literacy skills. The authors report that “the average respondent’s confidence in editing Wikipedia is relatively low” but that “about one in eight students had been given an assignment in

class at some point either to edit or create a new entry on Wikipedia” – which likely suggests that the (undisclosed by authors) university was one where at least one member of the faculty participated in the Wikipedia:Education Program. The vast majority (99%) of respondents reported having read an entry on Wikipedia, and over a quarter (28%) have had some experience editing it (interestingly, even when controlling for students who were assigned to edit Wikipedia, the former number is still as high as 20%).

Regarding the gender gap issues, women are much less likely to have contributed to Wikipedia than men (21% to 38%), and that becomes even more divergent when controlling for student assignments (13% to 32%). The authors find an indication of gender gap affecting the likelihood of Wikipedia’s contributions: students who are white, economically affluent, male and Internet-experienced are more likely to edit than others. The strongest and statistically significant predictor variables, however, are Internet skills and gender, and regression models show that variables such as race, ethnicity, socioeconomic status, time availability, Internet experience, and confidence in editing Wikipedia are not significant. The authors find that the gender becomes more significant as one’s digital literacy increases. At a low level of Internet skills, the likelihood of one’s contribution to Wikipedia is low, regardless of gender. As one’s skills increase, males became much more likely to contribute, but women fall behind. The authors find that women tend to have lower Internet skills than men, which helps explain a part of the Wikipedia gender gap: to contribute to Wikipedia, one needs to have a certain level of digital literacy, and the digital gap is reducing the number of women who have the required level of skills. The authors crucially admit that “why women, on average, report lower level understanding of Internet-related terms remains a puzzle. Although studies with detailed data about actual skills based on performance tests suggest no gender differences in the observed skills, research that looks at self-rated know-how consistently finds gender variation with real consequences for online behavior”. This suggests that while men and women have, in reality, similar skills, women are much less confident about them, which in turns makes them much less confident about contribut-

ing to (or trying to contribute to) Wikipedia. This, however, is a hypothesis to be confirmed by future research. In the end, the authors do feel confident enough to conclude that “gender and Internet skills likely have a relatively mild interaction with each other, reinforcing the gender gap at the high end of the Internet skills spectrum.” In conclusion, *this reviewer* finds this study to be a highly valuable one, both for the literature on gender gap and online communities, and for the Wikipedia community and WMF efforts to reduce this gap in our environment.

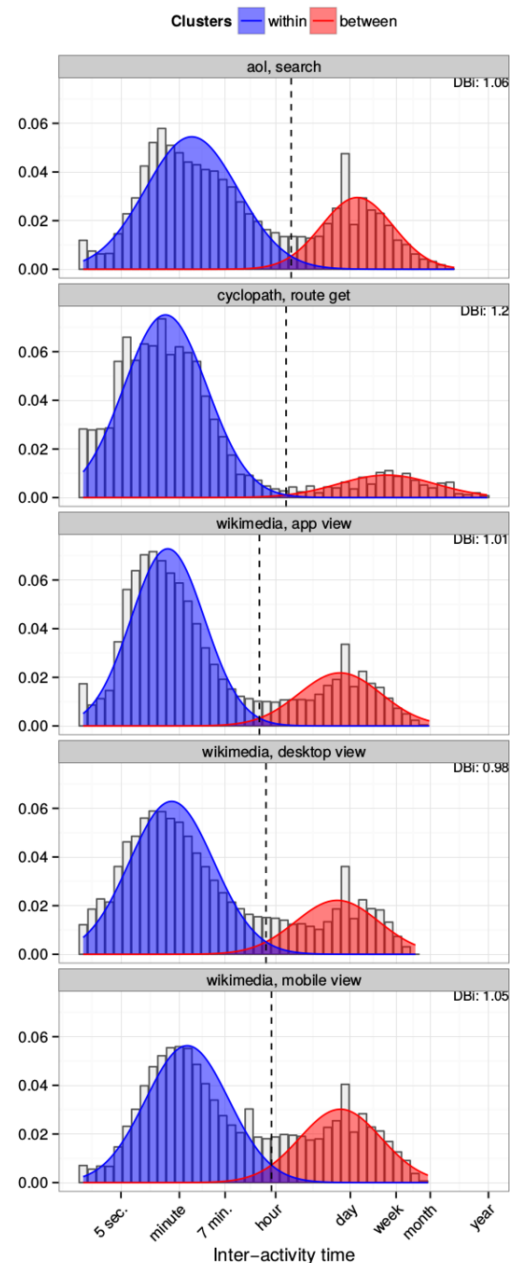
12.0.60 In nutritional articles, academic citations rise while news media citations decrease

A study published in *First Monday*^[2] analyzed the development of the referencing of 45 articles over nine topic groups related to health and nutrition over a period of five years (2007–2011) (unfortunately, the authors are not very clear on which particular articles were analyzed, and tend to use the concepts of an article and topic group in a rather confusing manner). Authors coded for references (3,029 total), information on editing history, and search ranking in Google, Bing and Yahoo! search engines. The study confirmed that Wikipedia articles are highly ranked by all search engines, with Yahoo! actually being even more “Wikipedia-friendly” than Google. The author shows that (as expected) the articles improve in quality (or at least, number and density of references) over time. Crucially, the authors show that the overall percentage of mainstream news media references has decreased, while references to academic publications increased over that time. By the end of the study period, only the article on (or topic group of?) *trans fat* contained more references to news sources than to academic publications. The authors overall support the description of Wikipedia as a source aiming for reliability, though they are hesitant to call it reliable, pointing out that for example 15% of analyzed references were coded as “outside the main reference type categories or... not be clearly determined”. The authors conclude, commendably, that “Wikipedia needs to be high on the agenda for health communication researchers and practitioners” and that “communications professionals in the health field need to be much more actively involved in ensuring that the content on Wikipedia is reliable and well-sourced with reliable references”.

12.0.61 Wikipedia user session timing compared with other online activities

reviewed by Maximilianklein (talk)

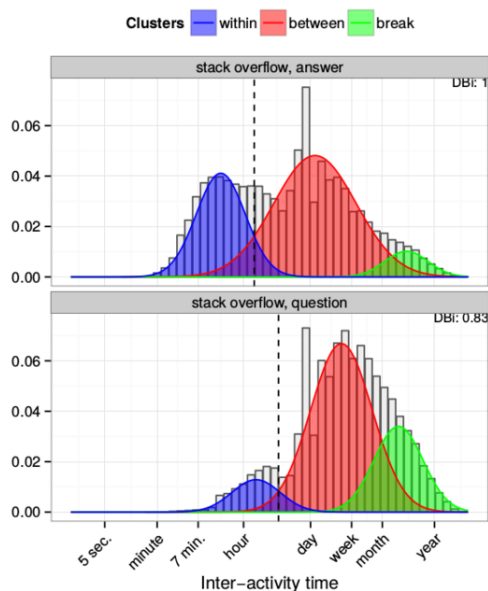
In a recent preprint titled “User Session Identification Based on Strong Regularities in Inter-activity Time”^[3],



Comparison of time between user interactions on Wikipedia, AOL and Cyclopath

Halfaker and team from the Wikimedia Foundation’s Analytics department and the GroupLens Lab ask whether there is some way we can talk about contributions in terms of “sessions” rather than atomic operations, in all collaborative work online. The researchers would like to answer “yes,” and that a “session” can be defined as the operations conducted until “a good rule-of-thumb inactivity threshold of about 1 hour” is reached, regardless if you’re editing Wikipedia, viewing Wikipedia, rating movies, searching AOL, or playing League of Legends. You may recall that Halfaker and Geiger came to a similar conclusion about “edit sessions” in a 2013 paper, but

now the idea is to cement that fact as a universal heuristic across many domains. Opposition to this idea has been that session length thresholds will always be arbitrary, or that a session deviates from completing a task that might extend beyond someone logging off for a night.



Stack Overflow user interactions

To bolster their argument, the authors use empirical data collected from seven datasets to test the hypothesis. The method employed is to take the log-normal time between user events, and then fit a bimodal distribution to the histogram. Once we have a two-humped histogram, we simply find the point which makes half the data “within” session and the other half “between” session.

AOL search data, Cyclopath route-getting requests, and Wikipedia viewing (from the desktop, mobile and apps) seem to fit bimodally. Together their the threshold is in the range of 29 to 115 minutes, but all would not be far off of an hour, say the authors. Yet when it comes to Wikipedia editing, OpenStreetMap editing, and MovieLens reviewing and searching, a bimodal 1-hour fit is good, but can be further explained by a trimodal model. In the case of the first two activities the third category is the wikibreak, and in the latter it is the ease the site make in rating movies in quick succession.

Even trimodally though, “this strategy for identifying session thresholds is not universally suitable for all user-initiated events”. For instance they show League of Legends, which has modal peaks at 5 minutes and one day. As a reviewer this is easy to describe from a player’s perspective. If you play 5 games in a row, which takes 5 minutes queuing between games, and then repeat it daily, you get the histogram seen where the 5 minute peak is about 5 times as tall as the day peak. Stack Overflow does not easily fit into their model at all with a threshold

of 335 minutes. The authors claim this is from the high quality edits expected at Stack Overflow.

Overall the authors conclude that one hour seems to suffice as a rule of thumb. But does it? The issue is that a goodness of fit with the bimodal models is not presented. This leaves outliers like Stack Overflow either able to be modeled but not compliant with the one hour rule, when they could just potentially not be describable using the proposed heuristic.

12.0.62 Briefly

“Wikimedia Movement in European countries as an example of civil participation”

This Polish-language book chapter^[4] (with an English abstract) looks at the Wikimedia community as a social movement. In the first subchapter, it argues that the Wikimedia movement is a type of new social movement which is fighting for equal access to free education. The bulk of subsequent subchapters consist of describing the European Wikimedia projects through tables listing whether they exist, estimated size in articles, members, etc., and briefly describing their activities such as involvement in the Wikipedia Loves Monuments initiative or with the GLAM sector. The book chapter is interesting as clearly placing itself in the relatively small body of literature that describe Wikipedia/Wikimedia as a social movement. Unfortunately it is primarily a descriptive rather than an analytical piece, and does not provide any significant theoretical justification for calling the Wikimedia movement a social movement, a weakness amplified by the fact that this work fails to engage with the prior relevant body of Wikipedia research, and is only very loosely connected to the literature on social movements.

Ranking public domain authors using Wikipedia data

This article^[5] proposes a way to combine Wikipedia and Online Books Page data, for the purpose of identifying the most notable (important, popular, read) authors whose work is about to enter the public domain, in order to facilitate and prioritize digitization of their works. The following information from the authors’ Wikipedia articles are used: “article length, article age in days, time elapsed since last revision, revision rate during article’s life, article text (200 topic weights derived from a topic model), category count, translation count, redirect count, estimated views per day, presence of translation for the 10 Wikipedias with the most translations, presence of bibliographic identifier (GND, ISNI, LCCN, VIAF), article quality classification (“Good Article” and “Featured Article”), presence of protected classification, indicator for decade of death for decades 1910–1950, and interactions between article age and all features.” The proposed al-

gorithm may be of interest to members of WikiProject Books, WikiProject Libraries, WikiProject Open, and related projects, as a means of generating an importance rating and selecting underdeveloped articles for development.

“Mining cross-cultural relations from Wikipedia - A study of 31 European food cultures”

The authors use^[6] Pierre Bourdieu's theories to analyze cultural similarities and differences between 31 European countries, by looking at the differences between articles on various national cuisines across 27 different European-language Wikipedias. They find that the existence, quality and links of studied Wikipedia articles can be correlated with data from the European Social Survey on cross-cultural ties between European countries. In addition to expected findings (all cultures are interested in their own cuisine first, then in famous ones such as French cuisine and in those of their neighbours), the article does present some interesting data, for example noting that the articles on Turkish cuisine are relatively well-developed on numerous Wikipedias, which could be explained by long-term and significant in size migration of Turkish people to various European countries, and the resulting interest in Turkish cuisine in those countries. The authors also find that significant differences do exist between different language Wikipedias, as different cuisines can be very differently described on different projects, thus reinforcing the theory that knowledge can be significantly influenced by one's culture. For Wikipedia editors, this is a reminder that all language editions suffer from significant biases, and that articles in different language editions can be and usually are significantly different.

Dissertation on automatic quality assessment

A recent PhD dissertation^[7] by Oliver Ferschke at the Technical University of Darmstadt “shows how natural language processing approaches can be used to assist information quality management on a massive scale” on Wikipedia. As the first main contribution, the author highlights his definition of a “comprehensive article quality model that aims to consolidate both the quality of writing and the quality criteria defined in multiple Wikipedia guidelines and policies into a single model. The model comprises 23 dimensions segmented into the four layers of intrinsic quality, contextual quality, writing quality and organizational quality.” Secondly, the dissertation presents methods for automatically detecting quality flaws (overlapping with previous publications co-authored by Ferschke), and evaluates them on a “novel corpus of Wikipedia articles with neutrality and style flaws”. Thirdly, the dissertation presents “an approach for automatically segmenting and tagging the user contributions on article Talk pages to improve work coordination among Wikipedians. These unstructured discussion

pages are not easy to navigate and information is likely to get lost over time in the discussion archives.”

39% of talk page threads contain wrong indentations

Ferschke's “English Wikipedia Discussions Corpus” (“EWDC”) is used in a paper^[8], to be presented at the 28th Pacific Asia Conference on Language, Information and Computing next month. In the paper, his doctoral adviser Irina Gurevych and another author construct a method to detect adjacency pairs (a user comment that responds to another) by analyzing the content, in particular detecting “lexical pairs” (giving the examples “(why, because)” and “(?, yes)”), validated against human annotation. As a side result, they observe that “Incorrect indentation (i.e., indentation that implies a reply-to relation with the wrong post) is quite common in longer discussions in the EWDC. In an analysis of 5 random threads longer than 10 turns each, shown in Table 1, we found that 29 of 74 total turns, or 39%±14pp of an average thread, had indentation that misidentified the turn to which they were a reply.”

Which talk page comment refers to which edit?

Another paper co-authored by Gurevych, titled “Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia”^[9] uses machine learning to automatically identify talk page comments about a particular article edit.

12.0.63 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- “Does the Administrator Community of Polish Wikipedia Shut out New Candidates Because of the Acquaintance Relation?”^[10] (cf. earlier coverage of related publications by the same authors: “Decline of adminship candidatures on Polish Wikipedia”, “What it takes to become an admin: Insights from the Polish Wikipedia”, “Predicting admin elections based on social network analysis”)
- “Development of a semantic data collection tool: The Wikidata Project as a step towards the semantic web.”^[11] (bachelor thesis)
- “To Use or Not to Use? The Credibility of Wikipedia”^[12]
- “Indexing and Analyzing Wikipedia’s Current Events Portal, the Daily News Summaries by the Crowd”^[13] From the abstract: “Wikipedia’s Current Events Portal (WCEP) is a special part of

Wikipedia that focuses on daily summaries of news events. ...First, we provide descriptive analysis of the collected news events. Second, we compare between the news summaries created by the WCEP crowd and the ones created by professional journalists on the same topics. Finally, we analyze the revision logs of news events over the past 7 years in order to characterize the WCEP crowd and their activities. The results show that WCEP has reached a stable state in terms of the volume of contributions as well as the size of its crowd...”



12.0.64 References

- [1] Hargittai, Eszter; Aaron Shaw (2014-11-04). “Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia”. *Information, Communication & Society* **0** (0): 1–19. doi:10.1080/1369118X.2014.957711. ISSN 1369-118X.
- [2] Messner, Marcus; Marcia W. DiStaso, Yan Jin, Shana Meganck, Scott Sherman, Sally Norton (2014-10-29). “Influencing public opinion from corn syrup to obesity: A longitudinal analysis of the references for nutritional entries on Wikipedia”. *First Monday* **19** (11). ISSN 1396-0466.
- [3] Halfaker, Aaron; Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, Morten Warncke-Wang (2014-11-11). “User Session Identification Based on Strong Regularities in Inter-activity Time”. *arXiv:1411.2878 [cs]*.
- [4] Patryk Korzeniecki: Ruch Wikimediów w państwach europejskich jako przykład aktywności obywatelskiej (Wikimedia Movement in European countries as an example of civil participation). Chapter 6 in: Joachim Osipiński, Joanna Zuzanna Popławska (eds.): *Oblicza społeczeństwa obywatelskiego*. WARSAW SCHOOL OF ECONOMICS PRESS, WARSAW 2014
- [5] Riddell, Allen B. (2014-11-08). “Public Domain Rank: Identifying Notable Individuals with the Wisdom of the Crowd”. *arXiv:1411.2180 [cs]*.
- [6] Laufer, Paul; Claudia Wagner, Fabian Flöck, Markus Strohmaier (2014-11-17). “Mining cross-cultural relations from Wikipedia - A study of 31 European food cultures”. *arXiv:1411.4484 [physics]*.
- [7] Ferschke, Oliver (2014-07-15). “The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia”. Darmstadt: Technische Universität Darmstadt.
- [8] Emily K. Jamison, Iryna Gurevych: Adjacency Pair Recognition in Wikipedia Discussions using Lexical Pairs. PDF
- [9] Johannes Daxenberger and Iryna Gurevych: Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia [<http://acl2014.org/acl2014/P14-2/pdf/P14-2031.pdf> PDF] Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 187–192, Baltimore, Maryland, USA, June 23-25 2014.
- [10] Sychała, Justyna; Mateusz Adamczyk, Piotr Turek (2014-06-30). “Does the Administrator Community of Polish Wikipedia Shut out New Candidates Because of the Acquaintance Relation?”. *International Journal On Advances in Intelligent Systems* **7** (1 and 2): 103–112. ISSN 1942-2679.
- [11] Ubah, Ifeanyichukwu (2013). *Development of a semantic data collection tool. : The Wikidata Project as a step towards the semantic web*.
- [12] Hilles, Stefanie (2014). “To Use or Not to Use? The Credibility of Wikipedia”. *Public Services Quarterly* **10** (3): 245–251. doi:10.1080/15228959.2014.931204. ISSN 1522-8959.
- [13] Tran, Giang Binh; Mohammad Alrifai (2014). “Indexing and Analyzing Wikipedia’s Current Events Portal, the Daily News Summaries by the Crowd” (PDF). *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. WWW Companion '14. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. pp. 511–516. doi:10.1145/2567948.2576942. ISBN 978-1-4503-2745-9. (ACM)

Wikimedia Research Newsletter

Vol: 4 • Issue: 11 • November 2014

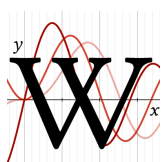
This newsletter is brought to you by the Wikimedia Research Committee and The Signpost


Subscribe:  [Email](#)  • [archives] [signpost edition] [contribute] [research index]

Chapter 13

Issue 4(12): December 2014

Wikimedia Research Newsletter



Vol: 4 • Issue: 12 • December 2014 [contribute]
[archives] 

Wikipedia in higher education; gender-driven talk page conflicts; disease forecasting

With contributions by: Federico Leva, Piotr Konieczny, Maximilian Klein, Tilman Bayer and Pine

13.0.65 Use of Wikipedia in higher education influenced by peer opinions and perception of Wikipedia's quality



The Universitat Oberta de Catalunya (Open University of Catalonia) in Barcelona, Spain

A paper titled “Factors that influence the teaching use of

Wikipedia in Higher Education”^[1] uses the technology acceptance model to shed light on faculty’s (of *Universitat Oberta de Catalunya*) views of Wikipedia as a teaching tool. The main factors are shown to be the perception of colleagues’ opinion about Wikipedia and the perceived quality of the information on Wikipedia. As the authors note, while prior studies also pointed to the quality concerns, this study suggests a causal link between colleagues’ views and one’s perception of Wikipedia quality. The authors conclude that the strong peer culture within academia makes the importance of role models very significant, which in turn has implications for the segment of the Wikimedia movement that desires greater ties with the academic world. The authors also note that “despite the lack of institutional support and acknowledgement, a growing number of academics think it is very useful and desirable to publish research results or even intermediate data in open repositories”, an attitude that also correlates positively with positive views of Wikipedia. To quote the authors’ very valid recommendation: “For those faculty members already using Wikipedia as a learning tool, we think it would have greater impact if they publicly acknowledged their practices more, especially to their close colleagues, and explain their own teaching experiences as well as the effects it has had on the students’ academic performance.” The team behind the paper is also partnering in the *Wikidata for research* project featured in *News and notes*.

13.0.66 Analysis of two gender-driven talk page conflicts on the German-language Wikipedia

Reviewed by Maximilianklein (talk)

“Gender differences within the German-language Wikipedia”^[2] is a pair of close readings of two gender-driven talk page conflicts on the German Wikipedia from 2006 and 2013, “show[ing] exemplarily that a) the feministic gender discourse in Wikipedia is not appreciated – primarily by male Wikipedians – [...] and b) that discussions behind the scenes of Wikipedia can feature an unpleasant and rude nature, that is not very appealing and motivating for female contributors”. The

analysis aims to focus on the communication styles of the gendered personalities as viewed under the critical rubrics of Margarete Jäger and Nina Schuppener. In the degenerating arguments around whether or not the welcome message on the German Wikipedia's main page (2006 thread) and German Wikipedia articles in general (2013/14 straw poll talk page) should use generic male pronouns and nouns, or newer more neutral alternatives, like using parentheses in "*Mitarbeiter(in)*", it is highlighted that the male-appearing participants use instruction and discrediting statements; and the female-appearing tend to question intellectual capabilities and give advice. Finally the authors conclude that "the most crucial point is the fact that the female author gave up [first]," stopping responding less than 24 hours into the discussion, and that the change advocated for was not enacted. These deconstructed examples add to an evidence of a hypothesis that minority voices are crowded out in Open Culture, as purported by the "Free as in Sexist" theory.

13.0.67 Briefly



"Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854" as seen in the English Wikipedia article *Epidemiology*.

History of the Spanish Wikipedia's ArbCom

A short recounting by Sefidari and Ortega (pre-print) summarised the history of the Spanish Wikipedia *Comité de resolución de conflictos* (arbitration committee), which existed from 2007 to 2008. It was composed of admins, received complaints which in 80 % of cases involved admins, dismissed nearly all cases presented, ruled against the claimant in a large majority of accepted cases, and was finally dissolved in 2009.^[3]

Two new papers on disease forecasting using Wikipedia

Yet another study (pre-print), considering 5 articles, showed that English Wikipedia page views trends can forecast the peak in influenza-like illnesses in the USA. Essentially, by visiting the articles in question, users are self-reporting their (suspect) disease, some weeks in advance of the data collected centrally by a government agency based on medical practitioners' reports of the same.^[4] Another study, again focused on some English Wikipedia articles, reached the same conclusion with slightly different (and, notably, fully open source) methods, for 14 diseases, while producing a useful list of some dozens past studies on the matter.^[5]

Wikipedia as a source of health information during salmonella outbreak

A statistically significant survey in the Netherlands assessed with what efficacy the population was informed about *Salmonella* infection during an outbreak in the country. Nearly all information was received passively (mainly from TV, radio and newspapers, but also social media); of the minuscule minority who actively sought information, most turned to their newspaper website, or ended up (with highest satisfaction among all sources) on official websites or Wikipedia.^[6]

Most MoodBar users became longer-term contributors

A study on one dataset produced by the (mostly discontinued) MoodBar tool showed that the newcomers who gave feedback via the MoodBar were significantly more likely to become longer-term contributors. After six months, 3.6% of editors who were able to use the MoodBar were still editing, compared to 3.3% of those who did not have the option.^[7]

New R libraries for Wikipedia research

A new R programming language library "*wikipedia-trend*"^[8] that facilitates longitudinal page-view analyses has been created. The package is a wrapper on top of long-time service `stats.grok.se/Wikipedia:Stats.grok.se/stats.grok.se`. This marks an uptick in the popularity of the R language for Wikipedia analysis as *WikipediR* was also recently released which itself wraps many common mediawiki API calls.

Use of Wikinews to teach journalism students

This paper^[9] discusses an educational project that used Wikinews in an undergraduate journalism course at the Australian University of Wollongong. While the use of

Wikipedia in education has dominated the relevant discussions, Wikinews seems like a valuable, yet underused tool for journalists-in-training. Though this essay-like paper seems to describe the experience in a positive fashion, it does not contain any specific conclusions, nor a list of articles edited by the students that would allow for a more-in depth commentary in the context of the Wikimedia learning experience.

“Linking Today’s Wikipedia and News from the Past”

This workshop paper^[10] presents a method to automatically identify articles in the New York Times archive matching a particular event mentioned on Wikipedia (dataset).

13.0.68 Other recent publications

A list of other recent publications that could not be covered in time for this issue – contributions are always welcome for reviewing or summarizing newly published research.

- **“An Empirical Study of Motivations for Content Contribution and Community Participation in Wikipedia”**^[11] From the abstract: “The research findings show that content contribution is more driven by extrinsically oriented motivations, including reciprocity and the need for self-development, while community participation is more driven by intrinsically oriented motivations, including altruism and a sense of belonging to the community.”
- **“Wikipedia as a Time Machine”**^[12] (presented at WWW 2014)
- **“Hacking Trademark Law for Collaborative Communities”**^[13] (related website: <http://collabmark.org/>)
- **“The political economy of wilkiality: a South African inquiry into knowledge and power on wikipedia”**^[14] (PhD Thesis)
- **“Predicting Low-Quality Wikipedia Articles Using User’s Judgements”**^[15] From the abstract: “In this paper, we utilize article ratings from Wikipedia users for the first time to assess article quality. We define ‘low-quality’ based on those ratings and design automatic methods to identify potential low-quality articles.”
- **“Infoboxer: Using Statistical and Semantic Knowledge to Help Create Wikipedia Infoboxes”**^[16]
- **“On the Use of Reliable-Negatives Selection. Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia.”**^[17]

13.0.69 References



- [1] Meseguer Artola, Antoni; Eduard Aibar Puentes; Josep Lladós Masllorens; Julià Minguillón Alfonso; Maura Lerga Felip (2014-12-11). “Factors that influence the teaching use of Wikipedia in Higher Education” (Article).
- [2] Sichler, Almut; Elizabeth Prommer (2014-12-22). “Gender differences within the German-language Wikipedia”. *ESSACHESS - Journal for Communication Studies* 7 (2(14)): 77–93. ISSN 1775-352X.
- [3] Sefidari, Maria; Felipe Ortega (2014-12-10). “Evaluating arbitration and conflict resolution mechanisms in the Spanish Wikipedia”. *arXiv:1412.3695 [cs]*.
- [4] Hickmann, Kyle S.; Geoffrey Fairchild; Reid Priedhorsky; Nicholas Generous; James M. Hyman; Alina Deshpande; Sara Y. Del Valle (2014-10-22). “Forecasting the 2013–2014 Influenza Season using Wikipedia”. *arXiv:1410.7716 [q-bio, stat]*.
- [5] Generous, Nicholas; Geoffrey Fairchild; Alina Deshpande; Sara Y. Del Valle; Reid Priedhorsky (2014-11-13). “Global Disease Monitoring and Forecasting with Wikipedia”. *PLoS Comput Biol* 10 (11). doi:10.1371/journal.pcbi.1003892.
- [6] Velsen, Lex van; DesiréJMA Beaujean; Julia EWC van Gemert-Pijnen; Jim E. van Steenbergen; Aura Timen (2014-01-31). “Public knowledge and preventive behavior during a large-scale Salmonella outbreak: results from an online survey in the Netherlands”. *BMC Public Health* 14 (1): 100. doi:10.1186/1471-2458-14-100. ISSN 1471-2458. PMID 24479614.
- [7] Ciampaglia, Giovanni Luca; Dario Taraborelli (2014-09-04). “MoodBar: Increasing new user retention in Wikipedia through lightweight socialization”. *arXiv:1409.1496 [physics]*.
- [8] Meissner, Peter. “Introduction to Public Attention Analytics with Wikipediatrend”. Retrieved 31 December 2014.
- [9] Blackall, David (2014). “Learning skills in journalistic skepticism while recognising whistleblowers” (PDF). The European Conference on Education 2014 Brighton, United Kingdom Official Conference Proceedings. Naka Ward, Nagoya, Aichi Japan: The International Academic Forum (IAFOR). ISSN 2188-1162.
- [10] Mishra, Arunav (2014). *Linking Today’s Wikipedia and News from the Past*. Proceedings of the 7th Workshop on Ph.D Students. PIKM '14. New York, NY, USA: ACM. pp. 1–8. doi:10.1145/2663714.2668048. ISBN 978-1-4503-1481-7. / preprint PDF
- [11] Xu, Bo; Dahui Li. “An Empirical Study of Motivations for Content Contribution and Community Participation in Wikipedia”. *Information & Management*. doi:10.1016/j.im.2014.12.003. ISSN 0378-7206.
- [12] Stewart Whiting, Joemon M. Jose, Omar Alonso: Wikipedia as a Time Machine. WWW’14 Companion, April 7–11, 2014, Seoul, Korea. PD

- [13] Welinder, Yana; Stephen LaPorte (2014-08-05). *Hacking Trademark Law for Collaborative Communities*. Rochester, NY: Social Science Research Network.
- [14] Ovesen, Håvard (2014). “The political economy of wilkiality: a South African inquiry into knowledge and power on wikipedia”.
- [15] Zhang, Ning; Lingyun Ruan; Luo Si (2015-01-01). “Predicting Low-Quality Wikipedia Articles Using User’s Judgements”. In Elisa Bertino, Sorin Adam Matei (eds.). *Roles, Trust, and Reputation in Social Media Knowledge Markets*. Computational Social Sciences. Springer International Publishing. pp. 91–99. ISBN 978-3-319-05467-4.
- [16] Roberto Yus, Varish Mulwad, Tim Finin, and Eduardo Mena: “Infoboxer: Using Statistical and Semantic Knowledge to Help Create Wikipedia Infoboxes” PDF
- [17] Edgardo Ferretti, Marcelo Errecalde, Maik Anderka, Benno Stein: On the Use of Reliable-Negatives Selection. Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia. In: Proceedings of the 25th International Workshop on Database and Expert Systems Applications (DEXA’14): 11th International Workshop on Text-based Information Retrieval (TIR’14), Munich, Germany, 2014. IEEE. PDF

Wikimedia Research Newsletter

Vol: 4 • Issue: 12 • December 2014

This newsletter is brought to you by the Wikimedia Research Committee and The Signpost

Subscribe:  **Email**  • [archives] [signpost edition]
[contribute] [research index]

13.1 Text and image sources, contributors, and licenses

13.1.1 Text

- **Research:Newsletter** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter?oldid=14441365> *Contributors:* Piotrus, Peteforsyth, Herbythyme, MZMcBride, Nemo bis, Rock drum, DarTar, Jodi.a.schneider, Trijnstel, Tbayer (WMF), Hexatekin, Suresh Rewar and Anonymous: 3
- **Research:Newsletter/2014/January** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/January?oldid=7442312> *Contributors:* Graham87, Tony1, Anomie, J Milburn, The ed17, Jtmorgan, Paine Ellsworth, DarTar, Ypnypn and Tbayer (WMF)
- **Research:Newsletter/2014/February** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/February?oldid=7856777> *Contributors:* Piotrus, Nemo bis, David Ludwig, The ed17, Wavelength, Nettrom, Citation bot, DarTar, Junkie.dolphin, Maximilianklein, Mr. Stradivarius, Tbayer (WMF), Jonesey95 and Halfak (WMF)
- **Research:Newsletter/2014/March** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/March?oldid=8022217> *Contributors:* Rich Farmbrough, Nikerabbit, John Vandenberg, Tony1, Nemo bis, Jim.henderson, Doc James, The ed17, Jtmorgan, Albany NY, Computermacgyver, DarTar, Tbayer (WMF) and Kimaus
- **Research:Newsletter/2014/April** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/April?oldid=10283235> *Contributors:* Tony1, Davidwr, The ed17, Cantons-de-l'Est, Junkie.dolphin, Mohamed CJ, Tbayer (WMF) and Jonesey95
- **Research:Newsletter/2014/May** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/May?oldid=11185224> *Contributors:* Graham87, Tony1, Nemo bis, DarTar, John of Reading, Wdchk, Maximilianklein, Tbayer (WMF), Jonesey95 and Anonymous: 1
- **Research:Newsletter/2014/June** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/June?oldid=9044597> *Contributors:* Graham87, Quiddity, Tony1, The ed17, Adler.fa, DarTar, Maximilianklein, Tbayer (WMF), Jonesey95, Quercus solaris, Kimaus and Anonymous: 1
- **Research:Newsletter/2014/July** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/July?oldid=9457363> *Contributors:* Piotrus, Graham87, Hanteng, The ed17, TomStar81, Qwfp, K6ka, DarTar, Maximilianklein, GoingBatty, Hfordsa, Tbayer (WMF), Sun Creator and Anonymous: 1
- **Research:Newsletter/2014/August** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/August?oldid=14435684> *Contributors:* Graham87, Tony1, Nemo bis, Wavelength, Cantons-de-l'Est, Qwfp, Elekhk, GregorB, Maximilianklein, Pine, Tbayer (WMF), Jonesey95, Greenrd and Dario (WMF)
- **Research:Newsletter/2014/September** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/September?oldid=14435703> *Contributors:* Piotrus, Graham87, Tony1, John Broughton, Puchku, Computermacgyver, Wdchk, Maximilianklein, GoingBatty, Pine, Tbayer (WMF), Jonesey95, Mark Schierbecker, Dario (WMF) and Anonymous: 1
- **Research:Newsletter/2014/October** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/October?oldid=10403331> *Contributors:* Jim.henderson, Cantons-de-l'Est, Maximilianklein, GoingBatty, Pine, Tbayer (WMF), Chris troutman, Kimaus and Dario (WMF)
- **Research:Newsletter/2014/November** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/November?oldid=10674351> *Contributors:* Piotrus, Graham87, Lambiam, John Broughton, Naddy, Swpb, Maximilianklein, Tbayer (WMF) and Dario (WMF)
- **Research:Newsletter/2014/December** *Source:* <https://meta.wikimedia.org/wiki/Research%3ANewsletter/2014/December?oldid=10942596> *Contributors:* Graham87, Nemo bis, Paine Ellsworth, Maximilianklein, GoingBatty, Pine, Tbayer (WMF), Daniel Mietchen, Jonesey95, EllenCT, AmericanLemming and Dario (WMF)

13.1.2 Images

- **File:ASUNCION_PANTEON_NACIONAL_DE_LOS_HÉROES.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/e/e/ASUNCION_PANTEON_NACIONAL_DE_LOS_H%C3%89ROES.jpg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Felipe Antonio
- **File:Ambrogio de Predis_-_Girl_with_Cheries_-_WGA18376.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/0/0b/Ambrogio_de_Predis_-_Girl_with_Cheries_-_WGA18376.jpg *License:* Public domain *Contributors:* Web Gallery of Art: `` `Image` `` *Info about artwork* *Original artist:* Giovanni Ambrogio de Predis (circa 1455-after 1508)
- **File:Article.edit_diff.lifetime.density.by_year.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/7/7f/Article.edit_diff.lifetime.density.by_year.svg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Halfak (WMF)
- **File:Belgium_provinces_regions_stripped.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/e/e5/Belgium_provinces_regions_stripped.png *License:* CC-BY-SA-3.0 *Contributors:* Merged from Image:Gemeenschappenkaart.png and Image:Belgium provinces blank.png by User:Stevenfruitsmaak. *Original artist:* User:Stevenfruitsmaak
- **File:Cscw2014wikiprojects_slides_ccbysa3.pdf** *Source:* https://upload.wikimedia.org/wikipedia/commons/5/5e/Cscw2014wikiprojects_slides_ccbysa3.pdf *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Jtmorgan

- **File:Feed-icon.svg** *Source:* <https://upload.wikimedia.org/wikipedia/commons/4/43/Feed-icon.svg> *License:* MPL 1.1 *Contributors:* feedicons.com *Original artist:* unnamed (Mozilla Foundation)
- **File:History_Wikipedia_English_SOPA_2012_Blackout2.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/a/a1/History_Wikipedia_English_SOPA_2012_Blackout2.jpg *License:* CC BY-SA 3.0 *Contributors:* Transferred from en.wikipedia to Commons by Sreejithk2000 using CommonsHelper. *Original artist:* Pseudoanonymous at English Wikipedia
- **File:Immanuel_Kant_(painted_portrait).jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/4/43/Immanuel_Kant_%28painted_portrait%29.jpg *License:* Public domain *Contributors:* /History/Carnegie/kant/portrait.html *Original artist:* unspecified
- **File:LatinroomHillman.jpg** *Source:* <https://upload.wikimedia.org/wikipedia/commons/e/e4/LatinroomHillman.jpg> *License:* CC BY-SA 3.0 *Contributors:* Own work by uploader, photo by Michael G. White *Original artist:* Crazypaco
- **File:New-Map-Sinophone_World.PNG** *Source:* https://upload.wikimedia.org/wikipedia/commons/f/fa/New-Map-Sinophone_World.PNG *License:* Public domain *Contributors:* Own work *Original artist:* ASDFGHJ
- **File:PLoS_Biology_cover_April_2009.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/9/9c/PLoS_Biology_cover_April_2009.svg *License:* CC BY 2.5 *Contributors:* Vol. 7(4) April 2009 PLoS Biology (direct link) *Original artist:* PLoS
- **File:Plutarchs_Lives_Vol_the_Third_1727.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/3/3c/Plutarchs_Lives_Vol_the_Third_1727.jpg *License:* Public domain *Contributors:* Private Collection of S. Whitehead *Original artist:* Plutarch, M. Dacier, Jacob Tonson, et al.
- **File:Rectorat_de_la_Universitat_Oberta_de_Catalunya.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/c/ca/Rectorat_de_la_Universitat_Oberta_de_Catalunya.jpg *License:* CC BY-SA 3.0 *Contributors:* Own work (own photo) *Original artist:* Pere López
- **File:SRL-Full-p40.png** *Source:* <https://upload.wikimedia.org/wikipedia/commons/0/0a/SRL-Full-p40.png> *License:* CC BY-SA 4.0 *Contributors:* Own work *Original artist:* Jasneet Sabharwal
- **File:Snow-cholera-map.jpg** *Source:* <https://upload.wikimedia.org/wikipedia/commons/c/c7/Snow-cholera-map.jpg> *License:* Public domain *Contributors:* ? *Original artist:* ?
- **File:Triangulation_translation_English_Hungarian_Romanian.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/a/ad/Triangulation_translation_English_Hungarian_Romanian.png *License:* CC BY 4.0 *Contributors:* http://www.lrec-conf.org/proceedings/lrec2014/pdf/864_Paper.pdf *Original artist:* Ács Judit
- **File:Twitter_logo_initial.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/8/8b/Twitter_logo_initial.svg *License:* Public domain *Contributors:* Own work, modified from File:Twitter logo.svg *Original artist:* Original uploader was en>User:GageSkidmore, modified by User:Cpro
- **File:Userssessions_stackoverflow.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/6/64/Userssessions_stackoverflow.png *License:* CC BY-SA 4.0 *Contributors:* <http://arxiv.org/abs/1411.2878> *Original artist:* Aaron Halfaker
- **File:Userssessions_wiki.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/5/5b/Userssessions_wiki.png *License:* CC BY-SA 4.0 *Contributors:* <http://arxiv.org/abs/1411.2878> *Original artist:* Aaron Halfaker
- **File:Vis_gartenbaukino3.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/f/fb/Vis_gartenbaukino3.jpg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Severin Dostal
- **File:WRN_2011.pdf** *Source:* https://upload.wikimedia.org/wikipedia/commons/a/a7/WRN_2011.pdf *License:* CC BY-SA 3.0 *Contributors:* m:Research:Newsletter, Vol.1 issues 1-6 *Original artist:* Text: Various contributors, see p.43 / Images: Various contributors, see p.44
- **File:WRN_2012.pdf** *Source:* https://upload.wikimedia.org/wikipedia/commons/d/d9/WRN_2012.pdf *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* DarTar
- **File:WRN_header.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/7/7d/WRN_header.png *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* DarTar
- **File:Waldo_Tobler_2007.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/b/b9/Waldo_Tobler_2007.jpg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Alvesgaspar
- **File:Wikimania_2012_Group_Photo-0001a.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/9/9a/Wikimania_2012_Group_Photo-0001a.jpg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Helpameout
- **File:Wikimedia_Research_&_Data_Showcase_-_February_2014.webm** *Source:* https://upload.wikimedia.org/wikipedia/commons/0/09/Wikimedia_Research_%26_Data_Showcase_-_February_2014.webm *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* DarTar
- **File:Wikimedia_Research_Newsletter.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/3/39/Wikimedia_Research_Newsletter.jpg *License:* CC BY-SA 3.0 *Contributors:* http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2011-07-25/Recent_research *Original artist:* Signpost contributors
- **File:Wikimedia_Research_Newsletter_Logo.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/c/cb/Wikimedia_Research_Newsletter_Logo.png *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* DarTar
- **File:WiktionaryEn.svg** *Source:* <https://upload.wikimedia.org/wikipedia/commons/f/ff/WiktionaryEn.svg> *License:* CC BY-SA 3.0 *Contributors:* ? *Original artist:* ?

13.1.3 Content license

- Creative Commons Attribution-Share Alike 3.0