Abstract

I here give an overview of Wikipedia and wiki research and tools. Well over 1,000 reports have been published in the field and there exist dedicated scientific meetings for Wikipedia research. It is not possible to give a complete review of all material published. This overview serves to describe some key areas of research.

1 Introduction

Wikipedia has attracted researchers from a wide range of disciplines—physicist, computer scientists, librarians, etc.—examining the online encyclopedia from a several different perspectives and using it in a variety of contexts.

Broadly, Wikipedia research falls into four categories:

1. Research that examines Wikipedia
2. Research that uses information from Wikipedia
3. Research that explores technical extensions to Wikipedia
4. Research that is using Wikipedia as a resource for communication

Research that examines Wikipedia look on how the encyclopedia evolves, how the users interact with each other, how much they contribute and most of this kind of research is not interested in the content per se. Vandalism in Wikipedia presents usually not a problem: It will just be one more aspect to investigate. Work in this area may benefit Wikipedia as it could help answering what rules the site should operate under, e.g., how beneficial is open editing with no user registration. One representative publication in this category is Jakob Voss 2005 article Measuring Wikipedia. One humorous quote is “Wikipedia cannot work in theory, but does in practice”: The major topic in this line of research is to find the answer to the question “why does it work at all?”

Research using information from Wikipedia will typically hope for the correctness and perhaps completeness of the information in Wikipedia. Many aspects of Wikipedia can be used, not just the raw text, but the links between components: Language links, categories and information in templates provide structured content that can be used in a variety of other applications, such as natural language processing and translation tools. Large-scale efforts extract structured information from Wikipedia and link the data and connect it as Linked Data. The papers with description of DBpedia represent examples on this line of research.

It is odd to write a Science 1.0 article about a Web 2.0 phenomenon: An interested researcher may already find good collaborative written articles about Wikipedia research on Wikipedia itself, see Table 1. These articles may have more complete and updated lists of published scientific work on Wikipedia, and much research-like reporting on Wikipedia of relatively good quality occurs outside ordinary academic channels, — on web-pages and blogs, and several non-academic organization have produced reports from large surveys, e.g., Pew Re-
Table 1: Wikimedia articles related to Wikipedia research. Some of these articles are in the main namespace, while others require the Wikipedia: namespace prefix, while others (m: prefixed) are on the meta wiki (meta.wikimedia.org).

<table>
<thead>
<tr>
<th>Wikipedia article</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>m:Research:Index</td>
<td>Primary entry point for Wikimedia research</td>
</tr>
<tr>
<td>en:Wikipedia</td>
<td>Main article about the encyclopedia</td>
</tr>
<tr>
<td>en:Criticism of Wikipedia</td>
<td>Long list of Wikipedia conference papers</td>
</tr>
<tr>
<td>en:Academic studies about Wikipedia</td>
<td>Comprehensive list of studies on Wikipedia</td>
</tr>
<tr>
<td>en:User:Moudy83/conference papers</td>
<td>Specific results on the growth of Wikipedia</td>
</tr>
<tr>
<td>en:Wikipedia:Modelling Wikipedia’s growth</td>
<td>Discusses quantitatively measures and links to various statistics</td>
</tr>
<tr>
<td>en:Wikipedia:Notability (academics)</td>
<td></td>
</tr>
<tr>
<td>en:Wikipedia:Survey_(disambiguation)</td>
<td>List of papers</td>
</tr>
<tr>
<td>en:Wikipedia:WikiProject Countering systemic bias</td>
<td>List resources for wiki research and researchers</td>
</tr>
<tr>
<td>en:Wikipedia:WikiProject Vandalism studies</td>
<td>Bibliography of scholar and science articles</td>
</tr>
<tr>
<td>m:Research</td>
<td></td>
</tr>
<tr>
<td>m:Wiki Research Bibliography</td>
<td>Overview of Wikipedia-related data</td>
</tr>
<tr>
<td>m:Wikimedia Foundation Research Goals</td>
<td></td>
</tr>
<tr>
<td>m:Research:data</td>
<td>Overview of research questions</td>
</tr>
<tr>
<td>en.wikiversity.org/wiki/Portal:Wikimedia Studies</td>
<td></td>
</tr>
<tr>
<td>s:Wikimedia-pedia</td>
<td></td>
</tr>
</tbody>
</table>

Wikipedia research continues to grow and now there are many thousands of research articles. In July 2010 Google Scholar claimed to return 196,000 articles when queried about Wikipedia, while Pubmed returns 55. In May 2012 I found 114 articles in PubMed, see also Figure 1 on page 1.

A few researchers have examined the development of research literature on Wikipedia through time. Han-Teng Liao reporting on his blog in 2012 found the number of theses from major Chinese-speaking regions to peak in 2009.5

Several researchers have already reviewed the Wikipedia and/or wiki research literature with varying degree of depth and breadth: A short 2009 report found 1’000 articles,7 and identifying 400 peer-reviewed articles a short 2009 paper summarized some of these articles.8 Another 2009 focused entire on the use of Wikipedia-derived data, e.g., for natural language processing and ontology building.9 On the other hand Nicolas Jullien’s 86-page long working paper10 focused on, e.g., editor motivation, process, contributor roles and quality. A 2012 version of the present paper11 together with a number of other initial papers by Montreal and Finnish research7,8,12,13 were evolved into a 138-page working paper.14 This lengthy paper was later split into several peer-reviewed journal papers.15,16

Why is Wikipedia research of interest?

Why does Wikipedia attract researchers? The popularity of the phenomenon probably attracts many, and most Wikipedia articles makes sense for the ‘common researcher’ in contrast to, say, bioinformatic databases that typically require ex-
The openness of the project and easy availability of data also makes Wikipedia of interest. Typically, research on the Web requires crawling many sites, whereas each complete language version of Wikipedia lies available for download as a compressed XML file ready for use and analysis. Other Web 2.0 large-scale data sets, e.g., from Facebook, may simply not be available for researchers, unless they get exceptional access to the data of the private company. The MediaWiki software also makes Wikipedia available available through the many-faceted API. The Toolserver facility, that was used by many Wikipedia programmers and some researchers, even enabled direct database queries. The Toolserver closed, but a similar service, Wikimedia Tool Labs, took over its functionality.

Multiple language editions make it possible to explore areas of language translation. Although the text of Wikipedia is not sentence aligned the large number of languages covered makes Wikipedia unique. With Wikidata multilingual labels for each concept can easily be obtained.

The availability of the revision history enables dynamic studies of content and contributors. In this aspect Wikipedia is not unique as studies of free and open source software based in public code repositories also make similar research possible.

The structured information in Wikipedia provided through categories and MediaWiki templates may also help researchers, and the sister project Wikidata provides an even more structured resource.

Researchers have also noted difficulties doing research with Wikipedia: As it is constantly changing a researcher may write about Wikipedia and later find it to be untrue, and the vast scale makes it difficult to do studies on the entire Wikipedia.19

## Tools and data sets

dumps.wikimedia.org provides compressed XML files of complete Wikipedias and their sister projects. This data contain the raw wiki markup along with metadata. Independent researchers have converted the wiki markup to an XML format, so that, e.g., links, lists and paragraphs are indicated.

### Table 2: Wikipedia data sets.

<table>
<thead>
<tr>
<th>WMF data sets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wikimedia Downloads</strong></td>
<td>Database backup dumps and static HTML dumps</td>
</tr>
<tr>
<td><strong>Page view statistics (‘raw’)</strong></td>
<td>Desktop raw page view statistics</td>
</tr>
<tr>
<td><strong>Page view statistics (‘all’)</strong></td>
<td>Full all page view statistics</td>
</tr>
<tr>
<td><strong>Repackaged page view statistics</strong></td>
<td>Page view statistics in a more compressed format.</td>
</tr>
<tr>
<td><strong>English Wikipedia pageviews by second 17</strong></td>
<td>Total page view statistics from the English Wikipedia by timestamp collected in March and April 2015</td>
</tr>
<tr>
<td><strong>Wikipedia Clickstream</strong></td>
<td>Referer-resource pairs from the request log</td>
</tr>
<tr>
<td><strong>Mediacounts</strong></td>
<td>Requests of media files on upload.wikimedia.org</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Third-party data sets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wikipedia XML Corpus 18</strong></td>
<td>Annotated in XML.</td>
</tr>
<tr>
<td><strong>DeepDive Open Datasets (WIKI)</strong></td>
<td>Natural language processing-annotated sentences from Wikipedia</td>
</tr>
<tr>
<td><strong>Scholarly article citations in Wikipedia</strong></td>
<td>Citation from Wikipedia articles to journal articles identified by PubMed, PubMed Central or DOI identifier.</td>
</tr>
<tr>
<td><strong>Structured citations in the English Wikipedia</strong></td>
<td>Citation metadata from the English Wikipedia dump.</td>
</tr>
<tr>
<td><strong>WikiCite data dumps</strong></td>
<td>Dump of the bibliographic data in Wikidata.</td>
</tr>
</tbody>
</table>
with XML tags rather than with wiki markup. They refer to this processed data as the “Wikipedia XML Corpus”, an one set of data from 2006 Wikipedias are available from http://www.connex.lip6.fr/~denoyer/wikiDataXML/

There are a number of other datasets, e.g., dumps.wikimedia.org/other/articlefeedback/ has article feedback data and https://frdata.wikimedia.org/ has data Wikimedia Foundation fundraising data. The MediaWiki API enables specialized queries, e.g., to get backlinks, blocks or use of templates, and to return the data in different formats such as XML and JSON.

Wikimedia makes usage statistics available to a certain extent, though full logs are not released in the open due to (the readers’) privacy. Simple page view counts are available from dumps.wikimedia.org/other/pagecounts-raw/. These statistics was originally collected by database engineer and former Wikimedia trustee Domas Mituzas (and distributed from dammit.lt/wikistats/). Further processing of this data is presented at http://stats.grok.se. JSON formatted data are available from stats.grok.se/json/. In September 2014 Andrew West discovered that these services underreported, — they did not count mobile views. Given the rise in mobile traffic the August 2014 the number of page views for the English Wikipedia view was underestimated considerable, — with about a third. New complete statistics are made available from https://dumps.wikimedia.org/other/pagecounts-all-sites/. This includes, apart from page views statistics to the default sites and to the mobile sites, also the Wikipedia Zero views.

With enabling a click tracking extension MediaWiki administrators can track users’ navigation around the wiki. During Wikipedia Usability Initiative Wikimedia enabled the extension for beta testing.

There are several derived data sets. The Wikilinks data set provides links from entities on webpages to English Wikipedia articles. It comprises around 40 million mentions and 3 million entities.

perlwikid, as the name implies, works with Perl and recent change patrolling program has been using it. Pwykediabot is a Python-based collection of tools for bot programming on Wikipedia and other MediaWikis. It can, e.g., create interlanguage links. WikiXRay is another software tool written in Python and R which may download and process data from the wikimedia site for generating graphics and data files with quantitative results.

Markus Krötzsch’s Java library Wikidata Toolkit can download and parse Wikidata dump files.

The command-line program wikipedia2text downloads a specified Wikipedia article and formats it for display on the command-line, while WikipediaFS makes raw text Wikipedia articles available under the Linux filesystem so a Wikipedia articles looks like an ‘ordinary’ file.

Edits on Wikipedias are relayed to a IRC server. Researcher may grab this real-time data, e.g.: In his dynamically updated webpage from 2011, Wikistream, Ed Summers, presents an aggregated continuously updated list with links to edited articles across the major language versions of Wikipedia.

The Wikimedia Toolserver was a platform for hosting various software tools written and used by Wikimedia editors. Programmers could apply for an account. It ran several specialized Web services, e.g., a user edit counter, and CatScan which enabled searching categories recursively. randombio of mzmcbride could sample randomly among bibliographies of living persons. See Table 3 on page 5 for other Web services. The Toolserver has now been replaced by Wikimedia Tool Labs and many of the Toolserver tools have been migrated there, e.g., catscan is now running from https://tools.wmflabs.org/catscan2/catscan2.php.

Research on ‘human subjects’

Several researchers have mentioned the problem of gaining access to users for surveys or interviews. The randomization tool on Wikipedia allows an unbiased sampling among registered editors. On the English Wikipedia the randomization tool has the address http://en.wikipedia.org/wiki/Special:Random/User. After sampling a set of editors researchers can contact the them by adding notes on the user talk pages or by emailing the user. However, editors may simply ignore the request, regard it as ‘survey spamming’, or simply have left the project. Emailing may also be hindered as the emailing facility in MediaWiki as users opt-in on the email contact ability.

Interview- or survey-based research may require the approval of the institutional review board as well as a recruitment form signed by the subject participating in the survey. Another issue raised in a blog post by Heather Ford in 2013 is whether to “onymize”, pseudonimize or anonymize, the research subjects. In Wikipedia research, as well as in Internet research generally, anonymization of a quote is difficult as the present day advanced Internet search engine usually has no problem tracking down text fragments. With the attribution inherent in the GPL and Creative Commons licences users may even insist on being attributed, — to quote Ford’s experience.
Table 3: The no longer functioning Toolserver and related Web services. See a large list at http://www.mediawiki.org/wiki/Toolserver/List_of_Tools.

<table>
<thead>
<tr>
<th>Name</th>
<th>Developer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article revision statistics</td>
<td>soxred93</td>
<td>Detailed overview of edits of a page, interactive graphs for edits over time, article size, and top page editors</td>
</tr>
<tr>
<td>Wikipedia page history statistics</td>
<td>aka</td>
<td>Detailed overview of edits of a page with graphs for edits over time, and page editors</td>
</tr>
<tr>
<td>X!'s Edit Counter</td>
<td>Several</td>
<td>Edit count for user with graphs over time</td>
</tr>
<tr>
<td>Edit Summary calculator</td>
<td>soxred93</td>
<td>Statistics about the edit summary with graphs</td>
</tr>
<tr>
<td>Contributors</td>
<td>Daniel Kinzler</td>
<td>Ranked list of contributors for a page</td>
</tr>
<tr>
<td>Recent change statistics</td>
<td>aka</td>
<td>Summarize recent edits (on German Wikipedia)</td>
</tr>
<tr>
<td>Revis</td>
<td>Finn Arup Nielsen</td>
<td>Sequential collaboration network visualization</td>
</tr>
<tr>
<td>Watcher</td>
<td>mzmcbride</td>
<td>Displays the number of users watching a page</td>
</tr>
<tr>
<td>Quarry</td>
<td>YuviPanda</td>
<td>Web-based SQL queries to Wikimedia databases</td>
</tr>
</tbody>
</table>

I had thought that this was the right thing to do: to anonymize the data, thus protecting the subjects. But the ‘subject’ was angry that he had been quoted ‘without attribution’. And he was right. If I was really interested in protecting the privacy of my subjects, why would I quote his sentence when anyone could probably Google it and find out who wrote it.

Books

A number of longer works describe Wikipedia and related topics from different aspects. New media journalist Andrew Lih’s book *The Wikipedia Revolution*\(^{25}\) recounts the history of Wikipedia and its controversies. *How Wikipedia works*\(^{26}\) by Wikipedians focuses on editing, policy and the community, while *MediaWiki*\(^{27}\) and *Working with MediaWiki*\(^{28}\) describes the MediaWiki software. The first book, *The Wiki Way*, presents the most central ideas of the wiki.\(^{29}\) *Critical Point of View: A Wikipedia Reader* is an edited book discussing a number of issues.\(^{30}\) There are several other books published.\(^{31–33}\)

Scientific meetings

Several dedicated scientific meeting centers around wikis and Wikipedia. The ACM affiliated meeting WikiSym presents results in all areas of wiki research. Since 2006 the *SemWiki* workshop has presented results in semantic wiki research. The research community around that meeting also interacts at the semanticweb.org site, — itself a semantic wiki. The *WikiAI* workshop concentrates on the interface between artificial intelligence with machine learning, computational linguistics on the one side and wiki and other collaborative-built knowledge bases on the other side. *Workshop on Collaboratively Constructed Semantic Resources* focuses on social semantics and social natural language processing with several contribution centered on the Wikipedia corpus. The 2009 workshop *The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* also featured Wikipedia research content.

The community meeting Wikimania focuses on Wikipedia and its sister Wikimedia foundation operated projects. Apart from community-related topics the meeting usually has a good deal of research-oriented material presented. The organizers publish no formal proceedings, but the Wikimania site has usually some description of the contributions, and Wikimedia Commons make videos from recent meetings available.

Other communication channels

WikiSym mailing list wiki-research has little activity, while Wikimedia mailing lists wiki-research-l and wiki-tech-l are fairly active.

Wikimedia Foundation-employed data analyst Erik Zachte makes charts, tables and comma-separated values files available from stats.wikimedia.org of large-scale analyses of all Wikimedia projects. His blog is available from www.infodisiac.com. The web service ‘Vital Signs’ running from Wikimedia Labs displays interactive charts of total pageviews and other metrics over all Wikimedia projects and all languages.

The newsletter *The Signpost* (Wikipedia Signpost) discusses different matters of Wikimedia projects. More critical is the forum
Wikipedia and individuals have composed sites with critical commentaries, see, e.g., www.wikipedia-watch.org. Wikipedia Weekly is a podcast with episodes from 2006 to presently 2009. Since July 2011 the meta-wiki of Wikimedia has published the monthly Wikimedia Research Newsletter (WRN) available from meta.wikimedia.org/wiki/Research:Newsletter. It focuses on new Wikimedia-related research. Wikimedia Foundation-employed Dario Taraborelli often writes entries, but individual researchers make substantial contributions. The newsletter is also aggregated into a single volume for each year. The 2012 edition was 95 pages long. The Twitter and Identi.ca user WikiResearch posts links to new research on Wikipedia.

2 Examining Wikipedia

Wikipedia has been examined in a number of ways, both quantitative and qualitative, and with an analysis of the full data set and just a small part.

What is Wikipedia?

The most basic question in the study of Wikipedia asks “what is Wikipedia”? Wikipedia itself reports its first sentence in October 2013 “Wikipedia is a collaboratively edited, multilingual, free Internet encyclopedia supported by the non-profit Wikimedia Foundation.” In the initial version of the Wikipedia article from November 2001 Larry Sanger defined Wikipedia as “the name of an open content, WikiWiki encyclopedia found at http://www.wikipedia.com/, as well as of its supporting, very active encyclopedia-building project.” Note that the recent quote defines Wikipedia as a work rather than as an organization, the difference between “free” and “open”, and between “collaboratively edited” and the more technical term “WikiWiki”. Wikidata presently (2013) claims it as an instance of a “wiki”, an “internet encyclopedia” and a “Wikimedia project” and the English description reads “free online encyclopedia that anyone can edit” while the German description use the word project. Interestingly, a study on the German Wikipedia community found that interviewees would see a German focus on the end product, while claiming the English-speaking community focused on the process citing a German Wikipedia: “We are not here because we want to use the wiki and have fun with it, but we want to have an encyclopedia which is bigger and better than any encyclopedia that has been there before [...]”

Whereas a definition of Wikipedia as an online encyclopedia would be from a reader’s point of view, regarding Wikipedia as a collaborative writing application (CWA) would also regard it from an editor’s point of view. In a review of health CWAs researchers would map CWA depending on use patterns: virtual communities (patients, e.g., Dealing with Autism), professional communities (e.g., RadiologyWiki) and Science 2.0 (e.g., OpenWetWare). The researchers would place Wikipedia at the very center of the map, see Figure 2.

Researchers have presented other definitions: Benkler and Nissenbaum referred to Wikipedia as a “project” and mentioned it as an example of commons-based peer production. Konieczny would call Wikipedia an “online organization [...]”, an “online community” and “related to several social movements, including the Free and Open Source Software Movement, Open Publishing Movement, and Free Culture Movement”. Such definitions focus not so much on the end product, but rather the process, the project and the group of people that led to the result.

Yet other researchers regard Wikipedia as a work beyond the encyclopedia, e.g., as a semi-structured knowledge source which can be used as a basis to derive explicit facts, or a “continuously edited globally contributed working draft of history”. Wikipedia might also be regarded as part of the Web 2.0 or social media.

The answer to the question “what is Wikipedia” directs the researcher in his/her study. If the researcher thinks Wikipedia is an encyclopedia then the researcher likely focuses on the content and most naturally compares Wikipedia against other reference works, — printed or online. If Wikipedia is a semi-structured knowledge source then Wikipedia should instead be compared to works such as WordNet and Semantic Web ontologies.

Quality

Several formal studies of the quality of Wikipedia have been performed. Such studies typically select a sample of Wikipedia articles and “manually” read and judge the quality, sometimes in comparison with other encyclopedias or other resources. The quality may be rated on several dimensions:

ii “Wikipedia” (oldid=576265305)
iii “Wikipedia” (oldid=331655554)
iv https://www.wikidata.org/wiki/Q52

Accuracy (no factual errors), coverage, bias, conciseness, readability, up-to-dateness, usable/suitable and whether the articles are well-illustrated and well-sourced. For Wikipedia’s ‘featured articles’ Wikipedia has the following quality dimensions: comprehensive, accurate, verifiable, stable, well-written, uncontroversial, compliance, appropriate images, appropriate style and focus, while Stvilia et al. in their study of Wikipedia discussions worked with 10 different dimensions.

There have been many quality studies in the health/medicine/drug domain: In a 2013 review of collaborative writing applications in health care researchers identified 25 papers reporting on the quality of information in such systems, with 24 of them evaluating Wikipedia, and researchers continues to study health-related information quality in Wikipedia.

**Overall quality**

In the perhaps most widely referenced investigation the science journalists of *Nature* collected 42 science articles from Wikipedia and *Encyclopædia Britannica* and let blinded experts evaluate them. The comparison between the two encyclopedias showed that the articles of Wikipedia contained the most factual errors, omissions and misleading statements, — but surprisingly not particularly many more than *Encyclopædia Britannica*: 162 against 123. Both encyclopedias contained 4 “serious errors, such as misinterpretations of of important concepts”. The Nature study was not itself peer-reviewed.

In the 2006-published study Roy Rosenzweig examined several quality aspects of American history articles in English Wikipedia compared against *Encarta* and American National Biography Online. He found the essays on the United States history to have inaccurate descriptions and with incomplete coverage. He attributed this to “broad synthetic writing is not easily done collaboratively” and found biographies of historical figures to “offer a more favorable terrain for Wikipedia since biography is always an area of popular historical interest”. Rosenzweig then examined bibliographies of American historical figures in *Encarta* and the 18,000 entries *American National Biography Online* for comparison against Wikipedia. Of 52 examined people listed in American National Biography Online about half were listed in Wikipedia and one-fifth in Encarta. He found the American National Biography Online had more details with about four times as many words as Wikipedia. He noted a bias in coverage between articles on Isaac Asimov, President Woodrow Wilson and Lyndon LaRouche judging the American National Biography Online to give a more proportionate coverage. He went on to examine factual errors. In 25 Wikipedia articles he found clear cut errors in four, and 3 articles with...
Table 4: Selection of Wikipedia quality studies. See also the Multimedia Appendix 3 of the Archambault 2013 review.³⁶

<table>
<thead>
<tr>
<th>Topic</th>
<th>Comparisons</th>
<th>Articles</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science²⁴</td>
<td>Britannica</td>
<td>42</td>
<td>Blinded experts</td>
</tr>
<tr>
<td>Biographies of American historical figures²⁰</td>
<td>Encarta, American National Biography Online</td>
<td>52/25</td>
<td>The author</td>
</tr>
<tr>
<td>Pop culture, current affairs, science³³</td>
<td>—</td>
<td>3</td>
<td>3 librarians</td>
</tr>
<tr>
<td>Surgical procedures⁴⁴,⁴⁵</td>
<td>—</td>
<td>30</td>
<td>Experts</td>
</tr>
<tr>
<td>General¹⁶,⁴⁷</td>
<td>Brockhaus</td>
<td>50</td>
<td>Research institute</td>
</tr>
<tr>
<td>Drug information</td>
<td>Medscape Drug Reference</td>
<td>80</td>
<td>The authors</td>
</tr>
<tr>
<td>Medical students information⁴⁹,⁵⁰</td>
<td>AccessMedicine, eMedicine, UpToDate</td>
<td>3</td>
<td>Blinded experts</td>
</tr>
<tr>
<td>Cancer information</td>
<td>US National Cancer Institute’s Physician Data Query (PDQ)</td>
<td>10</td>
<td>Medically trained personnel</td>
</tr>
<tr>
<td>Osteosarcoma⁵¹,⁵²</td>
<td>NCI patient and professional site</td>
<td>20</td>
<td>3 independent observers</td>
</tr>
<tr>
<td>Mental disorders</td>
<td>13 websites</td>
<td>10</td>
<td>Four drug information residency-trained pharmacists</td>
</tr>
<tr>
<td>Medication information</td>
<td>Manufacturer’s package insert</td>
<td>20</td>
<td>Four drug information residency-trained pharmacists</td>
</tr>
<tr>
<td>Orthognathic surgery</td>
<td>24 other websites</td>
<td>The topic</td>
<td>Scoring against “DISCERN”</td>
</tr>
<tr>
<td>Nephrology⁵⁷</td>
<td>—</td>
<td>95</td>
<td>Counting references, readability index computation</td>
</tr>
<tr>
<td>General⁵⁸</td>
<td>—</td>
<td>134</td>
<td>96 experts</td>
</tr>
<tr>
<td>Medical conditions</td>
<td>Peer-review literature</td>
<td>10</td>
<td>Pairs of medicine residents or rotating interns</td>
</tr>
<tr>
<td>Prescription drugs</td>
<td>—</td>
<td>22</td>
<td>The authors</td>
</tr>
<tr>
<td>Drugs⁵⁵</td>
<td>Text books</td>
<td>100</td>
<td>The authors(?)</td>
</tr>
<tr>
<td>Health, medicine, nutrition</td>
<td>WebMD, Mayo Clinic</td>
<td>92</td>
<td>Raters/authors</td>
</tr>
<tr>
<td>Breast reconstruction</td>
<td>9 web sites</td>
<td>1</td>
<td>Readability index computation, authors rating</td>
</tr>
</tbody>
</table>

factual errors among 10 examined for Encarta and 1 article with errors in American National Biography Online. Rosenzweig found Wikipedia “more anecdotal and colorful than professional history” and focus on topics with recent public controversy, and concluded “Wikipedia, then, beats Encarta but not American National Biography Online in coverage and roughly matches Encara in accuracy”, and further noted American National Biography Online has richer contextualization and easily outdistant Wikipedia on “persuasive analysis and interpretations, and clear and engaging prose”.

Another early quality of information study, a peer-reviewed one from 2006, looked on the credibility of Wikipedia and its articles.⁶⁸ However, this study was not a comparison and not blinded.

In December 2007 Stern performed an examination of the German Wikipedia and the on-line edition of the German encyclopedia Brockhaus.⁴⁶,⁴⁷ This weekly magasin had asked an independent research institute, Wissenschaftlicher Informationsdienst Köln, to evaluate 50 articles in relation to criteria of correctness, completeness, up-to-dateness and comprehensibility. In 43 cases the Wikipedia article was evaluated as better than Brockhaus’, and Wikipedia got the best grade average.⁶⁹,⁷⁰ Wikipedia was regarded as better in up-to-dateness and—perhaps surprisingly—in correctness, while Brockhaus scored better in completeness. Furthermore some Wikipedia articles were regarded as too
complicated for the lay reader.

In the summer of 2009 Danish newspaper Berlingske Tidende made a small informal comparison between the Danish Wikipedia and the larger (in terms of number of articles) expert-written Den Store Danske online encyclopedia. Overall the Danish Wikipedia came slightly ahead due to its many links, typically longer articles and more frequent updates, even considering the background of the authority of Den Store Danske. The author also noted the quicker and more precise searching in Wikipedia.

Among the many health-related quality of information studies is a study from 2007 where medical doctors reported on surgical information in Wikipedia. Identifying 39 common surgical procedures the researchers could find 35 corresponding Wikipedia articles with all of them judged to be without overt errors. The researchers could recommend 30 of the articles for patients (22 without reservations), but also found that 13 articles omitted risks associated with the surgical procedure.

Other researchers examined Wikipedia August 2009 cancer information and US National Cancer Institute’s Physician Data Query (PDQ). They found that Wikipedia had similar accuracy and depth when compared against the professionally-edited PDQ, but they also found that Wikipedia had lower readability as measured with the Flesch–Kincaid readability test.

Considering scope, completeness, and accuracy of information for osteosarcoma on April 2009 English Wikipedia compared against patient and professional sites of US National Cancer Institute (NCI) 3 independent observers scored the answers to 20 questions on a 3 point scale. Wikipedia scored lower compared to the two NCI versions, though the statistical test only showed significant difference against the NCI professional version.

Three psychologists with relevant expertise examined 10 topics in mental disorders across 14 websites with respect to accuracy, up-to-dateness, coverage, referencing and readability. Among the websites, beside Wikipedia, were NIMH, WebMD and Mayo Clinic and among the topics examined were “childhood onset of psychosis” and “gambling and depression”. Wikipedia scored high (“generally rated higher”) on accuracy, up-to-dateness and referencing, while low on readability.

While numerous studies have examined the quality of Wikipedia, one finds few fewer studies of the qualities of other wikis, e.g., the Wikipedia sister projects. In 2008 a lexicography study would claim that Wiktionary had a poor quality, while at the same time noting the comparative studies favor- able to Wikipedia. One of the few studies to compare Wiktionary with other language resources examined the German Wiktionary with GermaNet and OpenThesaurus and found that the scope of the three resources varied depending on which variations they look at, e.g., Wiktionary had the hightest number of word senses but the lowest number of synonyms.

Factual errors

Trusting specific facts on Wikipedia is questionable, as there might be typos, intentional or unintentional errors, biased presentation, or hoaxes, e.g., a misinformation on Wikipedia propagated to the obituary for composer Maurice Jarre on The Guardian web-site. As a research experiment a student had entered made-up Jarre quotes in Wikipedia immediately after Jarre’s death. An obituary writer working under a tight deadline picked up this information though it stayed in Wikipedia for only 25 hours. The hoax was only revealed after the student contacted the publishers. Similar vandalism that spreads to obituaries happened for Norman Wisdom.

Cautionary notes have been cast for the open wiki-model in cases where potentially hazardous procedures are described. Especially chemical and medical procedures and compounds may call for complete and accurate description. For medical drug information Kevin Clauson and his co-authors compared Wikipedia and Medscape Drug Reference (MDR), a free online “traditionally edited” database. They found that Wikipedia could answer fewer drug information questions, e.g., about dosage, contraindications and administration. In the evaluated sample Wikipedia had no factual errors but a higher rate of omissions compared to MDR. The authors could also find a marked improvement in the entries of Wikipedia over a just 90 days period. The study went on to mainstream media with headlines such as “Wikipedia often omits important drug information” and even “Why Wikipedia Is Wrong When It Comes To Prescription Medicine”. As noted by Wikipedians on a discussion page the study did not mention the fact, that one of the Wikipedia manual of styles explicitly encourages Wikipedia authors not to include dosage information with the present wording of “Do not include dose and titration information except when they are notable or necessary for the discussion in the article.” Thus in one of the 8 examined question categories the omissions on the

---

part of Wikipedia comes as an intention by consensus.

On a small comparison study on medical information with just 3 topics blinded experts found some factual errors in Wikipedia, — around the level of medical online resources UpToDate and eMedicine. AccessMedicine were found to have no factual errors among its 3 articles examined.49, 50

Coverage

One kind of critique often carried forth is that Wikipedia tends to have a emphasis on topics in pop culture, — the critique following the template “there are more entries for [a pop culture phenomenon] than for Shakespeare.” vii Is there a bias in the topical coverage of Wikipedia? Are there any other bias in coverage, e.g., with respect to gender and nationality?

Studies on topical coverage in Wikipedia often examine the number of Wikipedia articles within a given subject area and compare that number to associated numbers in works or databases from governments, well-established companies or other organizations, which then acts as a reference.79–84, 86 see Table 5. In 2005 Altmann could write that “[m]edical Informatics is not represented sufficiently since a number of important topics is missing”. He had compared the English Wikipedia to 57 terms he found in “Handbook of Medical Informatics”.79

Looking at outbound scientific citations in the English 2007 Wikipedia I found astronomy and astrophysics articles rather much cited compared to Journal Citation Reports from Thomson Scientific, but generally an overall agreement.80 Journal of Biological Chemistry got undercited but that changed after automated mass-insertion of genetic information.81 One peculiarity with the sample occurred for Australia botany journals. A Wikipedia project had produced a number of well-sourced articles on Banksia some reaching featured article status. The citation from these Wikipedia articles would skew the statistics.

By sampling 3000 articles from the English 2006 Wikipedia and categorizing them against the Library of Congress categories Halavais and Lackaff found categories such as social sciences, philosophy, medicine and law underrepresented in Wikipedia compared to statistics from Books in Print.82 The two latter categories had, however, on average a comparably large article size. They identified science, music, naval and, e.g., geography as over-represented, with music probably benefitting from fans contributions and other categories from the mass-insertion of material from public data sources such as United States Census. The two investigators could also find missing articles in the 2006 Wikipedia, when compared to three specialized encyclopedias in linguistics, poetry and physics. Halavais and Lackaff also noted some peculiarities in Wikipedia, e.g., extensive list of arms in the military category, comics fans to some extent driving the creation of articles in the fine art category and voluminous commentary on the Harry Potter series in the literature category.

For twentieth century philosophers Elvebakk compared Wikipedia against two online peer-review resources, The Stanford Encyclopedia of Philosophy and the Internet Encyclopedia of Philosophy, with respect to coverage of gender, nationality and discipline. She concluded that Wikipedia did not represent “the field of philosophy in a way that is fundamentally different from more traditional resources” in 2008.83 Wikipedia had far more articles about the philosophers than the two other resources and only some minor differences in fractions, such as a smaller fraction of German and French philosophers.

In a study on the efficiency of Web resources for identifying medical information for clinical questions Wikipedia failed to give an answer in a little above a third of the cases, while Web search engines, especially Google, were much more efficient. However, Wikipedia was more efficient than medical sites such as UpToDate and eMedicine in terms of failed searches and number of links visited, and the ‘end site’ that most often provided the ultimate answer from a Google search was Wikipedia.81 In another 2008 medical coverage study researchers found over 80% of ICD-9 and ICD-10 diagnostic codes in gastroenterology covered by Wikipedia.84 A similar study for nephrology found 70.5% of ICD-10 codes represented in August 2012.57 In another life science coverage study, researchers constructed a semi-automated program for matching LOINC database parts with Wikipedia articles. Of the 1705 parts they examined in October 2007 they found 1299 complete matches in Wikipedia with their semi-automated method, 15 partial matches and a further 15 matches from manual search, i.e., 1329 corresponding to 78%.86 They concluded that “Wikipedia contains a surprisingly large amount of scientific and medical data”.

A 2008 study compared the number of words in sets of Wikipedia articles with the year associated with the articles and found that articles associated with recent years tended to be longer, i.e., recency was to a certain extent a predictor for coverage: The length of year articles between 1900 and 2008

vi “Why are there more Wikipedia entries for Doctor Who than there are for Shakespeare?”80
<table>
<thead>
<tr>
<th>Topic</th>
<th>Comparisons</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical informatics(^{79})</td>
<td>Handbook of Medical Informatics</td>
<td>“Medical Informatics topics are not very well represented in the Wikipedia currently [2004/2005]”</td>
</tr>
<tr>
<td>Scientific citations(^{80,81})</td>
<td>Thomson Scientific Journal Citation Reports</td>
<td>Generally good correlation with scientific paper citations. Astronomy and banksia somewhat overcited. Dependence on bots.</td>
</tr>
<tr>
<td>General topics, physics, linguistics, poetry(^{82})</td>
<td>Library of Congress categories, Encyclopedia of Linguistics, New Princeton Encyclopedia of Poetry and Poetics, Encyclopedia of Physics</td>
<td>82% (physics), 79% (linguistics) and 63% (poetry) coverage</td>
</tr>
<tr>
<td>Twentieth century philosophers(^{83})</td>
<td>The Stanford Encyclopaedia of Philosophy and the Internet Encyclopaedia of Philosophy</td>
<td>Wikipedia had 534 philosophers covered while the other two had 60 and 49, respectively</td>
</tr>
<tr>
<td>Gastroenterology(^{84})</td>
<td>ICD-9 and ICD-10 codes</td>
<td>83% coverage</td>
</tr>
<tr>
<td>Philosophers(^{85})</td>
<td>Facts extracted from A History of Western Philosophy, A History of Western Philosophy, The Oxford Companion to Philosophy and The Columbia History of Western Philosophy</td>
<td>52% coverage</td>
</tr>
<tr>
<td>Medical terminology(^{86})</td>
<td>LOINC database</td>
<td>78% coverage</td>
</tr>
<tr>
<td>General topics</td>
<td>—</td>
<td>30% obtained ‘good’ or ‘excellent’ marks</td>
</tr>
<tr>
<td>US gubernatorial candidates and elections(^{87})</td>
<td>Number of real world candidates and elections</td>
<td>93% candidate coverage, 11–100% election coverage</td>
</tr>
<tr>
<td>Women(^{88})</td>
<td>National Women’s History Project</td>
<td>23/174 or 77/268 missing</td>
</tr>
<tr>
<td>Nephrology(^{57})</td>
<td>ICD-10 codes</td>
<td>70.5% coverage</td>
</tr>
<tr>
<td>Scientists(^{89})</td>
<td>Thompson Reuter list</td>
<td>22%–48% coverage</td>
</tr>
<tr>
<td>Drugs(^{65})</td>
<td>Pharmacology text books</td>
<td>83.8% (German), 93.1% (English)</td>
</tr>
</tbody>
</table>

Table 5: Selection of Wikipedia coverage studies.

and the year as a predictor variable had a Spearman correlation coefficient on 0.79. The results were not homogeneous as the length associated with articles for Time’s person of the year had a correlation of zero with the year. Academy award winning films and “artist with #1 song” had correlation between the two: 0.47 and 0.30, respectively. The authors of the study also examined other sets of articles in Wikipedia and the correlation with column inches in Micropaedia of the Encyclopaedia Britannica, country population and company revenue. The correlations were 0.26, 0.55 and 0.49, respectively. In their comparison with 100 articles from Micropaedia they found that 14 of them had no Wikipedia entry, e.g., “Russian Association of Proletariat”, “League for the Independence of Vietnam” and “urethane”.\(^{92}\)

Bill Wedemeyer presented the quality of scientific articles on the English Wikipedia on Wikimania 2008 as he and his students had examined the coverage based on several data sets. On a cross-section of 446 articles randomly and blindly sampled from Encyclopaedia Britannica Wikipedia articles lacked entries for 15, e.g., “Bushmann’s carnival”, “Samarkand rug” and “Catherine East”. All of 192 random geographical articles from Britannica had corresponding articles in Wikipedia. Of 800 core scientific topics selected from biochemistry and cell biology text books 799 could be found in Wikipedia. He concluded that science is better covered than general topics and that Wikipedia covers nearly all encyclopedic topics.\(^{93}\)

Kittur, Chi and Suh developed an algorithm that would assign a topic distribution over the top-level categories to each Wikipedia article.\(^{94}\) After evaluating the algorithm on a human labeled data set they examined the English Wikipedia and found that ‘Culture and the arts’ and ‘People and self’ as the most represented categories. Between the 2006 and 2008 data set they found that ‘Natural and physical sciences’ and ‘Culture and the arts’ categories grew the most. By combining the algorithm
with a method for determining degree of conflict of each article, they could determine that ‘Religion’ and ‘Philosophy’ stood out as the most contentious topics.

A case of bias in coverage with an individual Wikipedia article reached mainstream media. A user flagged the article on Kate Middleton’s wedding grown for deletion. The flagging and the ensuing debate about the notability of the dress was seen a symptom of the ‘gender gap’ of Wikipedia. Jimmy Wales argued with a “strong keep” that “I believe that our systemic bias caused by being a predominantly male geek community is worth some reflection in this context” and pointed out that Wikipedia in contrast has “over 100 articles on different Linux distributions, some of them quite obscure” and with “virtually no impact on the broader culture, but we think that’s perfectly fine.” Parallel to the media focus on the gender imbalance among contributors, a couple of studies have examined on the possible biased representation of women on Wikipedia. Reagle and Rhue have reported on the female proportion in biographic resources for persons born after 1909: 28.7% (Gale Biographical Resource Center) and 24.5% (Wilson’s Current Biography Illustrated), but also as low as 15% (American National Biography Online). Other lists of notable persons yield percentage on 10% (The Atlantic top 100 most influential figures in American history) and 12% (Chambers Biographical Dictionary). For the English Wikipedia the researchers found 16% and after a similar analysis of Encyclopædia Britannica the researchers concluded “Wikipedia and Britannica roughly follow the biases of existing works”. In 2011 Gregory Kohs would report a higher number on 19% for the female proportion, — this was for a random sample of 500 living people biographed on Wikipedia. Reagle and Rhue also compared biographic article lengths in Wikipedia and Encyclopædia Britannica with respect to gender and found no consistent bias in either female or male direction. Wikidata, where a property of an entity may indicate the gender, can scale up the analysis and make it multilingual to several hundred thousand persons, even over a million: Using Wikidata’s ‘sex’/’gender’ property and its language links Max Klein compared the sex ratio across Wikipedia language versions finding (among the big Wikipedias) the most equal rate on the Chinese Wikipedia yet still well below 25% female, while the English Wikipedia had a female percentage on 17.55%. A reference file, the Virtual International Authority File (VIAF) data, gave a 24.35% female rate in Klein’s study. In 2015 Max Klein published a blog post with a more detailed reporting of the results of his study performed together with Piotr Konieczny. Among variables in relation to gender they considered celebrity status finding that “recorded females [in Wikipedia] are more likely to be celebrities”.

The studies find an increasing ratio of female representation over time (e.g., date of birth). Why does this increase appear? Is it because women have increased their position in society? Han-Teng Liao put forward an alternative ‘male gaze’ hypothesis, where the increase comes through a gender-interest bias, e.g., young males interested in female celebrities such as porn actresses.

Tools associated with Wikidata can make coverage estimation across Wikipedias trivial. The Mix’n’match lists entries from several external databases and display matches with Wikidata items, e.g., it can list the members of the European Parliament based on information from its homepage http://www.europarl.europa.eu/meps/together with the Wikidata item. Statistics can then show that (in the case with this database) all members are matched to Wikidata items, but, e.g., only 293 members, corresponding to around 8%, have a Danish Wikipedia article, while the English Wikipedia has a coverage of over 50% with 2020 articles for parliament members (as of October 2014). The European Parliament list is only one of several. Examples of other catalogues are Oxford Dictionary of National Biography, BBC Your Paintings and Hamburgische Biografie.

Wikipedians with a particular interest in coverage organize themselves in the WikiProject Missing encyclopedic articles, a project where the main goal is “to ensure that Wikipedia has a corresponding article for every article in every other general purpose encyclopedia available”. The project lists quite a number of reference works and other resources for benchmarking Wikipedia coverage.

https://commons.wikimedia.org/wiki/Category:Jane_Darnell.
amples are Encyclopædia Britannica 1911, 2007 Macropædia and The Encyclopedia of Robberies, Heists and Capers.

Editor and researcher of Wikipedia, Emilio J. Rodríguez-Posada, has attempted to estimate the number of “notable articles needed to cover all human knowledge”. As of January 2014 the number stands on 96 million. It is made up of, e.g., the number of species, in one source estimated to be 8.7 million (eukaryotes). The around 14 million entities on Wikidata (as of January 2014) makes up around 15% of 96 million, yet the 96 million does not include, e.g., the majority of the large number of chemical compounds described. Chemical Abstracts Service announced in 2011 that they had reached the 60 millionth entry in their registry.

Limitations in coverage due to Wikipedia policy of notability has inspired other web-sites: Deletionpedia records deleted pages from Wikipedia in a MediaWiki run site with no editing possible, and Obscuropedia is a wiki for non-notable topics not covered by Wikipedia.

**Up-to-dateness**

Several quality comparison studies examine the up-to-dateness and find that Wikipedia compares well in this aspect, although not unequivocal. In the comparison between Wikipedia and Medscape the researchers found four factual errors in Medscape among 80 articles examined. Two of these occurred due to lack of timely updates. No factual errors occurred in Wikipedia. The Wedemeyer study found that Wikipedia was much better up to date than Encyclopædia Britannica.

In a study on twentieth century philosophers Wikipedia had far more articles on philosophers born after the Second World War than two other online encyclopedias The Stanford Encyclopedia of Philosophy and The Internet Encyclopedia of Philosophy.

The Danish Wikipedia has a large number of bibliographies copied more or less unedited from two old reference works with expired copyright: Dansk biografisk Leksikon and Salmossens Konversationsleksikon. The age of the works affects the language and viewpoint of the Wikipedia articles.

Information on the death of a TV host, Tim Russert, came out in Wikipedia before news organizations published it. The author of the Wikipedia entry came inside a news organization.

Medical drugs may have associated safety alerts, e.g., United States Food and Drug Administration issues Drug Safety Communications with warnings and recommendations on medical drug use. When FDA issues these communications Wikipedians should incorporate them in the articles about the drug for timely information to patients and physicians. A study on these communications from 2011 and 2012 showed that Wikipedians do not update the information satisfactorily: For 22 prescription drug articles researcher found that Wikipedians had not incorporated specific FDA communications in 36% of the articles when they examined the articles more than a year after the FDA communications.

**Sources and links**

Many studies and tools examine the outbound references to sources that Wikipedia uses or the inbound links that comes from documents to Wikipedia. Often the count of sources are used as a feature in studies of article quality. Wedemeyer’s study looked on the references of Wikipedia articles. They found that most developed articles had sufficient references comparable to a scientific review article, but some articles, even two featured, had insufficient referencing.

Ed Summers’ Linkypedia Web service available from http://linkypedia.info/ makes statistics available online on which Wikipedia articles links to specific webpages on selected websites. As of 2012 statistics was mainly available for selected GLAMs. It enable, e.g., British Museum to see that their page on ancient Greece had the highest number of inlinks from Wikipedia: 39 as of February 2012; that Wikipedia articles in the category “BBC Radio 4 programmes” linked much to their website; that the “Hexne Hoard” article had no less than 27 links to their website; and that the total number of Wikipedia links to the British Museum website was 2,673 from 1,209 pages.

Another Web service of Ed Summers, wikitweets, displays microposts from Twitter that link to Wikipedia. The Web service runs in real-time from wikitweets.herokuapp.com and tweets with excerpt of the linked Wikipedia articles are stored in machine readable JSON format at the Internet Archive (archive.org/details/wikitweets). Readers clicks the links in Wikipedia articles to outside sources to a considerable extent. In 2014 a CrossRef statistics reported that Wikipedia was the “8th largest referrer of CrossRef DOIs”. Web services from CrossRef allow easy identification of which Wikipedia articles cite a scientific article based on DOI information as well as a real-time citation events updates.

---

How Wikipedia is used as a source has also been described. In the media business, e.g., Philadelphia Inquirer instructs journalist never to use Wikipedia “to verify facts or to augment information in a story” and one reporter has been cited for “there is no way for me to verify the information without fact-checking, in which case it isn’t really saving me any time.” Other news organizations allow occasionally citation of Wikipedia as a source, e.g., Los Angeles Times. An analysis of press mentioning (cited, quoted or referred) of early Wikipedia found, e.g., that Daily Telegraph Online accounted for roughly a third of all citations. The site consistently referred to Wikipedia for further reading and background information in sidebars.

Genre and style

Researchers have also investigated other content topics besides quality. In one study researchers examined 15 Wikipedia articles and their corresponding talk page and compared them with other online knowledge resource: Everything2 and Columbia Encyclopedia. They specifically looked on the formality of the language by counting words indicative of formality or informality, such as contraction, personal pronouns and common noun-formative suffixes. With factor analysis they found that the style Wikipedia articles is close to that of the Columbia Encyclopedia.

Lexical analysis was featured in a study on Wikipedia biased representation wrt. gender. The study found that, e.g., the word ‘divorce’ appear with a higher rate in articles about women compared to articles about men.

The genre may also evolve as editors extends and change the articles.

Accessibility

Lopes and Carriço examined 100 Wikipedia and 265 non-Wikipedia Web articles cited by Wikipedia. They looked for their level of accessibility, i.e., to which extent the fulfilled the Web Content Accessibility Guidelines of the World Wide Web Consortium designed “to make Web content accessible to people with disabilities”. The authors found that Wikipedia articles on average scored better than the Web articles they cited. They further argued that the discrepancy between the accessibilities could lower the credibility of Wikipedia. It is not so odd that Wikipedia scores well in accessibility since HTML mark is automatically contructed from wiki-markup, and the software can be programed to ensure that, e.g., the ‘alt’ field of the ‘img’ HTML tag is automatically set.

Use of Wikipedia in court

The supreme court of India used Wikipedia for the definition of the word ‘laptop’, and several American courts have used Wikipedia in their rulings. e.g., Connecticut Supreme Court cited Wikipedia for the number of gay Congressmen. These cases are not singular: In February 2007 Washington Post noted that courts cited Wikipedia four times as often as Encyclopædia Britannica and in 2008 Murley ran a database search and found “1516 articles and 223 opinions that had cited to Wikipedia articles” in the Westlaw’s and ALLCASES databases. The English Wikipedia maintains incomplete lists of mostly English language court cases using Wikipedia as a source: Wikipedia:Wikipedia as a court source and Wikipedia:Wikipedia in judicial opinions. A couple of other language versions of Wikipedia have similar lists for cases in their respective languages.

Three lengthy papers examine the judiciary use of Wikipedia and discuss the controversy of using Wikipedia as an authority. They find the first references to appear in 2004, peaking in 2007 and a decrease towards the end of their examined periods, — November 2007 for Breinholt and 2008 for Stoddard. Breinholt classifies the different uses of Wikipedia into four categories:

1. Wikipedia as a dictionary. Wikipedia used, e.g., to answer what “candystriper” means.
2. Wikipedia as a source of evidence. In the most perilous use of Wikipedia judges rely on Wikipedia for evidence with one example being a judge relying on Wikipedia for whether Interstate-20 US highway does or does not extend from California.
4. Judiciary commentary about Wikipedia. One among the few cases involved a judge cautioning against citing Wikipedia in an appellant brief.

In some cases Wikipedia may be the only reference available for definitions of words, e.g., at one point Google returned only Wikipedia for a ‘define’ query on “candystriper”. Should one entirely ignore Wikipedia? I my opinion Wikipedia articles can be used for definitions provided that the definition has been overlooked by many readers and
editors and that many reliable editors have over time edited the article, — so we may regard it as a consensus definition. For establishing that ‘many reliable editors have edited’ one would need to examine the revision history and possibly the discussion page and its associated revision page. This process might require an expert Wikipedian.

Cases in Wikipedia and Wikimedia Commons may give rise to legal discussions. The copyright status around the so-called Monkey selfie has probably been the the most widely discussed. The essay ‘Final exam for wikilawyers’ by the Wikipedian Newyorkbrad sets up number of interesting questions from fictional and real-world cases.

Size across languages

Why does the language editions of Wikipedia differ in size? If it is often pointed out that the number of speakers of a language is a good indicator for the size of Wikipedia in that language, then why is the Norwegian Wikipedia larger than the Danish? And why was the Esperanto larger than the Arabic until 2011?

Morten Rask analyzed 11 Wikipedia language editions with respect to creation date of Wikipedia, number of speakers of the language, Human Development Index, Internet users, Wikipedia contributors and edits per article and found a number of correlations between these variables, e.g., the Internet penetration and level of human development was correlated to the number of contributors. Wikipedia contributors are rich and e-dy. Explaining the quality of the German Wikipedia Sue Gardner put forth related factors, saying: “Germany is a wealthy country. People are well educated. People have good broadband access. So the conditions for editing Wikipedia are there.” Other variables that may affect the Wikipedia size of different language edition are culture of volunteer, willingness to translate (from other language Wikipedia) and problems with non-latin characters. Among the reasons for the relatively small size of the Korean and Chinese Wikipedias Shim and Yang suggested the competition faced by Wikipedia from other knowledge-sharing Web services: Korean question/answering site Jisik iN and Chinese online encyclopedia Baidu Baike. As a further factor Andrew Lih and Sue Gardner would also mention the ability to meet face-to-face due to German-speakers geographically location in a relatively small area and the German Verein culture.

The Arabic Wikipedia has had a relatively small size compared to the number of speakers. The low attendance for a Wikipedia event in Egypt was blamed on ‘general lack of awareness of the importance of the issue’ and ‘culture of volunteer work’. Arabic users may choose to write in English because they find it easier to communicate in that language due to keyboard compatibility problems and to bring their words to a wider audience. Users of the Internet may also be hindered by low cable capacity in some areas as has been the case in East Africa.

One obvious factor for the size of a Wikipedia comes from the willingness of the community to let bots automatically create articles. Letting bots create articles on each species may generate many hundreds of thousands articles. The unwillingness to let bots roam and the elimination of stub articles and a focus on quality compared to quantity in the German Wikipedia may explain the why the its article-to-speaker ratio is (as of February 2014) quite lower than the Dutch and Swedish.

Network analysis, matrix factorizations and other operations

Modern network analysis has come up with a number of new notions, e.g., small world networks, the power law of scale-free networks, PageRank and hubs and authority. Formula with algorithms have been put forward that quantatively characterize the concepts and they have been applied to a diverse set of networks, e.g., the network of movie actors, power grid, neural network and the world wide web. Wikipedia researchers have also examined the quantitative characteristics for the networks inherent in Wikipedia. Among the many networks characteristics reported for a variety of Wikipedia derived networks are PageRank and Kleinberg’s HITS or other eigenvalue-base measures, small world coefficient, with cluster-coefficient and average shortest path, the size of the ‘bow tie’ or giant components, power law coefficients, $h$-index, reciprocity, assortativity coefficients, triade significance profil and acceleration.

Networks can be represented in matrices, thus matrices can also be constructed from content and metadata in Wikipedia articles. Mathematical operations can be performed on the matrices to examine aspects of Wikipedia or to test computational algorithms on large-scale data. Wray Buntine built a matrix from the within-wiki links between 500’000 pages of the English 2005 Wikipedia and used a discrete version of the hubs and au-
authority algorithm to find topics in Wikipedia, e.g., one topic would display the Wikipedia articles “Scientific classification” and “Animal” as the top authorities and “Arterial hypertension” and “List of biology topics” as the top hubs. Another approach builds a normalized matrix from the adjacency matrix of the within-wiki links between the articles. By augmenting the normalized adjacency matrix with an extra term the so-called Google matrix can be formed. The first eigenvector associated with the Google matrix determines the PageRank of an article. The adjacency matrix may be transposed, normalized and augmented. Its first eigenvector may be found to yield what has been called the CheiRank, — a concept closely linked to Kleinberg’s idea of hubs in his HITS algorithm and the analysis of Buntine. PageRank and “CheiRank” may be combined to form what has been called the 2DRank. In the family of these analysis Bellomi and Bonato wound in an analysis from 2005 find, e.g., United States, France, Russia, World War II, Jesus and George W. Bush on the top of the HITS authority and PageRank lists. Longitudinal analysis of these measures have also been performed.

Barcelona Media researchers performed wikilink network analyses of only the biographic articles of Wikipedia, — but across 15 different language versions. They found that PageRank for the English Wikipedia put America presidents highest on the ranked list. Comparing betweenness centrality across language versions showed overlaps among the top five most central persons with American Presidents often in the top five, as well as singers and World War II figures. However, cross-cultural differences also existed, e.g., with Chinese leaders central in the Chinese Wikipedia, Latin American revolutionaries central in the Spanish and Catalan Wikipedias, and Pope John Paul II central in the Polish and Italian Wikipedia. A gender gap existed as only three women appeared among the top five central persons across the 15 languages: Elizabeth II, Marilyn Monroe and Margaret Thatcher.

Yet another wikilink analysis restricted the analysis to Western philosophers—only analyzing 330 pages and 3706 links. They used the Prefuse toolkit for network visualization and also employed graph simplification under the name “Strongest LinkPaths”. With the influence network constructed from “influenced by” and “influenced” fields of the infoboxes on the philosophers Wikipedia pages they showed that Kant had influenced the largest number of other philosophers listed in Wikipedia.

Instead of working from the links, the words of a Wikipedia articles may also be used as features in the construction of a matrix, so the resulting matrix is a document-term matrix. A decomposition of such a matrix is often termed latent semantic analysis, particularly if singular value decomposition is the decomposition method. For assessing the performance of newly developed algorithms Řehůrek constructed a document-term matrix from the entire English Wikipedia with the resulting size of 100,000 times 3,199,665 corresponding to a truncated vocabulary on 100,000 words and almost 3.2 million Wikipedia articles. Analyses of document-term matrices can form the bases for content-based filtering methods and recommender systems, e.g., a system applied Latent Dirichlet Allocation on the 4 635 sized 2007 Wikipedia Selection for Schools dataset, potentially providing an alternative method for school children to navigate this corpus.

The category network forms another dataset. Chris Harrison has made large “clusterball” visualizations of the three levels of Wikipedia category pages. Another group of his Wikipedia visualizations, WikiViz, display millions of vertices.

Several researchers have considered the graph-/matrix formed when editors co-edits articles, i.e., the articles-by-editors matrix or co-authorship network. In an 2006 analysis of the articles-by-editors matrix constructed from of the Danish Wikipedia using non-negative matrix factorization on the (12774 × 3149)-sized matrix I found co-edited pattern on: Danish municipalities, Nordic mythology, Danish governments, article discussions, churches, science computers, pop culture, years, Scania, sport, countries and cities, plants and animals. As the edits are distributed unequal with respect to number of edits, both over articles and users, the results from the matrix factorization depended on the normalization of the matrix. With an unsuitable normalization identified co-edit cluster may simply reflect the contribution of single prolific author. With suitable normalization the factorization shows that articles and editors may be characterized by their co-editing patterns. Another work on the article-by-editor network used an algorithm to identify maximal bicliques, showing that some of the dense clusters of articles and editors were related to controversial topics (Scientology, creationism) or projects (WikiProjectElements for elements in the periodic table). Laniado et al. reported a large number of network metrics for the co-authorship network over time and topics after the main contributors for each article had been identified with a method by Adler et al.,
e.g., even after the thinning of the co-authorship network they found that over 96% of the contributors are connected in the giant component and with removal of 5000 contributors with the highest betweenness centrality the authors 85.9% author were still connected in the component. The analysis of the co-editing pattern has also been combined with geographic resolving of editors using natural language processing of user pages.\footnote{156}

A \textit{sequential collaboration network} is a network where users are represented as nodes and links between one editor and another are present if the second editor edits an article right after the first editor edited the article.\footnote{157} The network is a directed graph and may be said to be a social network, as the adjacency matrix is a \textit{user-by-user} matrix, albeit the editing may only be a form form of indirect collaboration. Some researchers call it the “article trajectory”.\footnote{158} Several different patterns might oc- currence when the sequential collaboration network is visualized for different articles. Iba et al. name “snake”, “wheel” and “star” as “egoboooster” networks.\footnote{159} An example of a sequential collaboration network is displayed in Figure 3 for an article of a Danish Politician on the Danish Wikipedia. Other applications of sequential collaboration networks have appeared for an online service on multiple business-related articles that adds color-coding based on sentiment analysis.\footnote{159,160}

When networks are formed from the \textit{reply network} of article discussion/talk pages or the activity on the user talk pages (\textit{user talk network} and \textit{wall network}) the networks better represent the ‘real’ social interaction among users in contrast to indirect stigmergic collaboration seen in the network formed from articles co-authorship. Laniado et al. construct such networks and identify the presence of \textit{welcomers}, — the users and bots writing a welcome message on new users’ talk pages, as well as reports on the user associativity. They also show that long chains of replies between users on article talk pages appear for articles such as “Intelligence design”, “Gaza War” and “Brack Obama”.\footnote{141}

\textbf{Article feedback}

In September 2010 Wikimedia launched an experimen- tal article feedback tool letting users evaluate articles as well-sourced, neutral, complete or readable on a five-point scale. Results of an analysis presented by Howie Fung of 1,470 ratings across 289 articles showed that registered users tended to give lower ratings than unregistered users. The maximum score came as the most frequent given score among unregistered users on all four dimensional. Registered users chose the minimum score as the most frequent on the dimensions of well-sourcedness and completeness, while they rated neutrality and readability higher.

\textbf{Hoax content}

Bloggers have made experiments with Wikipedia adding factual errors to observed how fast they get correct — if at all.\footnote{161,162} How long the hoax survive may depend on how obvious it is. P.D. Magnus made vandalisms anonymously and from various IP addresses across separate philosophy articles, e.g., the edit ‘Kant’s poetry was much admired, and handwritten manuscripts circulated among his friends and associates’ in the ‘Immanuel Kant’ article. After 48 hours 18 of 36 of this kind of vandalism remained.\footnote{163,164} Danish computer science student Jens Roland had his short article about the fictitious municipality \textit{Æblerød} surviving 20 months on the Danish Wikipedia and translated into several other language versions.\footnote{165} The scam was supported by external web-sites with information that looked genuine. For the cases of the bibliographies on John Seigenthaler\footnote{166} and Bertrand Meyer\footnote{167} that drew much media attention the vandal- ism remained for 132 and four days, respectively. In both these cases the false statements made could possible have let to criminal charges against the author. The English Wikipedia maintains a list of Wikipedia hoaxes (the page \textit{Wikipedia:List of hoaxes on Wikipedia}) showing that complete hoax articles may persist for many years.

Perhaps the most elaborate single incident con- cerned the article for the fictitious “Bicholim con- flict”, which persisted for well over 5 years. It was even rewarded with “Good Article” status, casting serious doubts on the article reviewing process on the English Wikipedia. This article was sup- ported by citations to fictitious books.\footnote{168} These kinds of hoaxes resemble the kind you also find in ordinary reference works, e.g., “Dag Henrik Erum-Hellerup” — a fictitious entry in a music reference work.

T. Mills Kelly teaches the course \textit{Lying About the Past} at George Mason University where students fabricate histories and add them to Wikipedia. In 2008 the tale about Edward Owens survived un- til Kelly announced the hoax at the end of the semester, while a 2012 experiment on the fabricated history of serial killer Joe Scafe survived 26 minutes after an announcement on Reddit.\footnote{169} Most of the hoaxes seem to arise from humor or critical examination. When prolific and barnstar-awarded editors insert false information it may re- quire a major cleanup, such as in the case with user Legolas2186. On the English Wikipedia he
Figure 3: Sequential collaboration network\textsuperscript{157} for Danish politician (Ellen Trane Nørby) on the Danish Wikipedia generated with my revvis online tool on the Toolserver. The size of the nodes are determined by the number of edits of the user. The article begins with the 213 IP address creating the article (upper left corner of plot) followed by an edit by the 212 IP address. At one point a Danish parliament IP (194.255.119.1) edits several times.

was blocked indefinitely after inserting false statements that was supported by citations to invented articles.\textsuperscript{170} His case seems not to originate in humor or critique. Should such a case be explained as a kind of addiction to barnstars and edit counts?

Vandalism

For obvious vandalism, where large part of an article is deleted, a 2004 study showed that it typically only took a couple of minutes before an article gets reconstructed.\textsuperscript{171} With data up to 2006 Reid Priedhorsky and his co-authors found that 5% “damaged” edits among 57.6 million revisions 42% of article damages were reverted within one estimated page view.\textsuperscript{172} Scientists have observed a lower rate of vandalism on specialized scientific topics: The edits on Wikipedia articles related to the WikiProject RNA and the Rfam database reach a vandalism rate of slightly over 1% (actually “possible vandalism”).\textsuperscript{173}

In his ethnographic study Lorenzen observed the Wikipedia:Vandalism in Progress page in two time periods in October and November 2005 to see how vandals were dealt with.\textsuperscript{174} During the time periods he noticed hundreds of reports of vandalism, while not all were reported vandalism were in fact vandalism. He reported 16 false reports and 39 user bans in the time period. He also discussed the issue of subtle vandalism escaping detection.

Among the temporal aspects examined in a 2006 study is the number of reverts through time from 2002 to the beginning of 2006. The researchers found that the number of reverts, including fast reverts have risen almost monotonically from below 1% to over 6%, and “that may signal an increasing amount of vandalism per page”. An exception on the monotonicity was at the introduction of the 3-revert rule established in November 2004 which almost halved the so-called double-reverts.\textsuperscript{175}

The communal ‘recent change patrol’ watch over the recent changes in Wikipedia on an entirely voluntary basis and edits or deletes vandalism. Wikipedians have conducted many tools, with names such as WikiGuard and WikiMonitor, for monitoring and semi-automated editing and
vandalism reversion. On http://www.wpcvn.com Dmitry Chichkov constructs a dynamic web page that displays an overview of recent changes along with extra useful information such as editors ‘karma’. More advanced tools does reversion automatically. An example is ClueBot. The early tools were mostly rule-based applying simple heuristics, but vandalism detection may also be viewed as a machine learning problem where the task is to classify an edit as a vandalism or not. Geiger and Ribes give a narrative of the reverting and blocking of a vandal using a combination of assisted editing tools and bots showing how multiple Wikipedia editors as vandal fighters use Huggle and Twinkle and how the tools interact together with ClueBot over a fifteen minute period. They also show the role of user talk pages in signaling warning messages to the user and to other vandal fighters and their tools. The presently operating ClueBot NG bot automatically reverts much simple vandalism very fast. Semi-automated anti-vandalism tool STiki have been reported to have a median age of reversion on approximately 4.25 hours.

Dmitry Chichkov has also reported descriptive statistics on the most reverted articles. He has made statistics sorted on revert ratio and filtered for number of revisions larger than one thousand available from wpcvn.com/enwiki-20100130.most.reverted.txt. Sexual and vulgar articles tend to get high on the list, e.g., 69, Nipple, Urine, Butt and Pussy, but also Geometry, Bubonic plague and Italian Renaissance for some reason.

Based on categories by Viégas et al. Priedhorsky et al. worked with the following 7 categories (features) of Wikipedia damage/vandalism: Misinformation, mass delete, partial delete, offensive, spam, nonsense and other. Having three independent judges and using a majority vote procedure they found 53% of damages with the nonsense feature, 28% with the offensive feature, 20% misinformation, 14% partial delete and the rest of the features occurring in less than 10% of the damages. In their discussion they noted that mass deletion is easy to detect while the more often occurring feature misinformation “may be the most pernicious form of damage”.

Biased editing

In biased editing editors diverge from the neutral point of view of Wikipedia, — one of its cornerstones. It may be associated with conflict of interest. Whereas graffiti vandalism is relatively easy to spot, biased editing comes usually from competent editors and biased edits may be harder to judge. To boost their reputation organizations or individuals may make such edits or they may pay others, e.g. public relations companies, to do it: Either deleting negative information or add positive. Law Professor Eric Goldman predicted in a 2005 blog post that in five years Wikipedia would fail due to “gamers” and “marketeers”. In 2013 Wikipedia is still fairly neutral, but there has been several cases of biased edits.

One early reported case from 2007, that exemplifies the problem, had Microsoft offering to pay blogger Rick Jelliffe to change Wikipedia articles on open-source document standards and the Microsoft rival format. The company was sure that the articles contained inaccuracies. Controversial edits from an IP address associated with podcasting entrepreneur Adam Curry in the “podcasting” Wikipedia article have also been noted. The investigation of biased edits became considerably easier when American PhD Student Virgil Griffith created and in August 2007 opened the web-based tool called Wikiscanner. It would merge IP address information with Wikipedia anonymous edit information. The tool pinpointed edits from organizations, which were exposed in mainstream media. Some of these were harmless spelling correction, others graffiti vandalism, yet others removal of negative information. Examples of the latter category are: Edits from the Vatican on Sinn Fein leader Gerry Adams removed links to news stories that alleged a link to murders in 1971 and edits from voting machine supplier Diebold removed unfavorable information about political funding and alleged rigging of the 2000 United States election. BBC edits rewrote a part on political correctness of the ‘Criticism of the BBC’ article. The Wikiscanner has even revealed edits from the public relations firm Hill & Knowlton to the Politics of the Maldives Wikipedia articles. The firm was employed in 2004 by the former President to improve the image of the country. The edits removed words such as “torture”, “propaganda” and “censorship”. After exposure of Wikipedia articles the Dutch justice ministry would temporarily block its 30.000 employers from using Wikipedia at work. A Wikiscanner, constructed by Peter Brodersen, is also available for the Danish Wikipedia. An extension of Wikiscanner using trademark and Ip2location databases would compute link distance between pages and categories and were reported to detect “the majority of COI edits”. Biased edit stories continues to pop up in mainstream media from time to time. In 2013, e.g., a user from City of Melbourne edited in the Occupy
Melbourne article, and the staff of Danish politicians were found to edit Danish Wikipedia articles on their politicians. As troublesome Wikipedia content Andrew Lih would also point to university articles with “glowing accolades, trumpeting selective accomplishments in the opening paragraphs as the most significant, noteworthy events, even if taken from obscure sources.”

In 2011 a French court ruled that a company had to pay its competitor company damages after someone from the former company had edited the Micropaiement article on Wikipedia.

A user may create multiple accounts, and if used for deception Wikipedia editors speak of ‘sockpuppets’. By their multiplicity sockpuppet accounts can create a sense of consensus, bias polls and work around Wikipedia’s three-revert-rule for fighting other editors, — an instance of so-called ‘gaming the system’. Through text analysis, looking for repeated spelling errors and idiosyncracies in sentence construction, socket puppets may be identified. One sockpuppet was exposed after another user examined the irrelevant references in an article about a company of questionable notability. In the ensuing deletion discussion, where several other “users” defended the company article, sufficient suspicion arose to warrant a sockpuppet investigation, that would eventually uncover a network of 232 confirmed sockpuppet accounts. These accounts predominantly edited in articles about companies and persons being promotional in nature.

The Wikiganda, a student project mentored by Virgil Griffith and at one point available from www.wikiwatcher.com, made use of automated text analysis to detect biased edits. The opinion mining and sentiment analysis technique used a lexicon of over 20,000 words from General Inquirer and Wiebe wordlists so each revision could get a ‘Propaganda Score’ and labeled as negative, positive or ‘vague’ propaganda. Also using the WikiTrust system and evaluating against 200 manually labeled revisions the system showed a precision/recall performance of 52%/63%. Our similar online system focused on business-related Wikipedia edits.

A special case of biased editing happened with Gibraltar and high profile Wikipedian Roger Bamkin. Bamkin did not change content per se, but rather promoted Gibraltar-related content in the ‘Did You Know’ section of the main page.

Wikipedia has the lengthy “Conflict of interest editing on Wikipedia” article, and a Facebook group called Corporate Representatives for Ethical Wikipedia Engagement (CREWE), established in the beginning of 2012, points to articles and members discuss the issues around conflict of interest edits. Read also the description of the Facebook group by DiStaso and the Wikipedia article about the Facebook group. Chartered Institute of Public Relations has established best practice guidelines for public relation professions, and David King from a firm offering Wikipedia consulting argues for “transparent community collaboration”.

Several companies offer paid editing, with one even claiming to have Wikipedia administrators among the editors that can help customers with changing Wikipedia. The journalist Simon Owen gained contact with one of their clients, who said that “they paid between $500 and $1,000 to have the page created, then an additional $50 a month afterwards for ‘monitoring’.”

Does the demographics of Wikipedia contributors—the prototype being the young male procrastinating graduate students—bias the content of Wikipedia? In 2013 American writer Amanda Filipacchi and editor Elissa Schappell noticed that the English Wikipedia subcategorized the “American Novelists” category into “American Women Novelists” but not “American Male Novelists”. Filipachi would call this “Wikipedia’s Sexism”, and one could hold the view that relegating women to a subcategory would lower the visibility of female writers. The case would be noted by several media outlets, and lengthy discussions ensued on Wikipedia. Apart from the discussion of sexism, Wikipedia contributors also discussed more generally about subcategorization, particularly about subcategorization based on ethnicity, gender, religion and sexuality. Before the “American Women Novelists” case there had, e.g., been discussions about the category “Male actors by nationality” and its subcategories in 2012.

The result of the discussion was to keep the subcategorization, and as of July 2013, e.g., the “Danish male actors” (of November 2012) persists but has been applied far less widely than “Danish actresses” (also of November 2012). Wikipedia maintains guidelines for categorization by ethnicity, gender, religion, or sexuality and a version reads: “Do not create categories that are a cross-section of a topic with an ethnicity, gender, religion, or sexual orientation, unless these characteristics...”
are relevant to the topic.” and “In almost all cases, gendered/ethnic/sexuality/religion-based categories should be non-diffusing, meaning that membership in the category should not remove membership from the non-gendered/non-ethnic/etc parent category.” The end of the American novelists subcategorization controversy is as of July 2013 that no individual pages are grouped there and writers are all moved to either “American women novelist” or “American male novelist”.

**Authorship**

A page will almost always have multiple authors. The revision history records each author contributions, but the format of the revision history makes it not trivial to determine who contributed what and the most, since text may be reformulated, moved, deleted and reintroduced.

To get an overview of the edits the convenient program history flow takes the revision history as input and visualizes the entire history of an article with colorings determined by author. Another related tool—WikiDashboard—also generates a visualization of the edit activity of each Wikipedia page. It embeds the generated plot in a proxy copy of the Wikipedia article showing the amount of edits of each author through time for the given article. Yet others have presented tools with examples of color coding of text depending on authorship—so-called blaming tools. Author color coding (‘Einfärben’) operates in.full in the German competing encyclopedia Wikiweise. Yet another tool, WikiBlame, helps in identifying versions of a Wikipedia article where a phrase occurs. The MediaWiki extension Contributors just counts the number of edits.

These approaches do not reveal if editors just copied text from other articles in Wikipedia. The detection presents a challenge that other researchers also ignore. For a full examination of the originality of an entry one would need to go through the entire revision history up to the date of the entry of all articles in Wikipedia. The same problem arises with translations: An entry may be novel across language versions or ‘just’ a translation.

A ‘good’ deletion, e.g., of vandalism or poorly written entries, increases the value of a Wikipedia article, so to which extent should a deletion count in the rank of authorships? In user reputation modeling deletion by administrators are considered to represent good deletions. Authorship determination has importance for redistribution of Wikipedia: Editors release their Wikipedia contributions under a share-alike license and this license requires that everyone copying or distributing Wikipedia lists ‘at least five of the principal authors of the Document’. A strict interpretation of this requirement call for a list of Wikipedia editors, e.g., in copies of Wikipedia articles on other web-sites. A quick look makes it evident that redistribution web-sites do not usually honour this strict interpretation.

Dynamic social network analysis may be used to examine editor contribution and role. Network visualization may identify “egoboosters” that relentlessly checks and edits an article.

**User contributions**

Wikipedia authors contribute with varying amount of work. A relatively few number of editors make the great percentage of the number of edits, e.g., Jimmy Wales reported that ‘half the edits by logged in users belong to just 2.5% of logged in users’ based on a count from December 2004. A study based on a 2007 dump of Wikipedia found that 46% of registered users made only one edit. In 2008 Ortega et al. would quantify the contributor inequality for registered users with the Gini coefficient across several Wikipedias: It was between 0.924 and 0.964.

Are the many edits of elite editors of major value? Perhaps the elite editors are just fixing, e.g., minor comma errors. Based on a small investigation of a few Wikipedia articles Aaron Swartz argued that the occasional contributor is actually the one that make substantive edits. Swartz had looked on the “Alan Alda” article and results from “several more randomly-selected articles” showed the same pattern. A Minnesota 2007 study considered the “impact of an edit” taking the number of page views into account with a metric the authors called persistent word view (PWV), and after some discussions on the Wikipedia research mailing list in November 2008 Alain Desilets summarized with the following assertions:

- Most edits done by a small core
- But, most of the text created by the long tail
- However, most of the text that people actually read, was created by the small core

---


\( ^{xxi} \)http://lists.wikimedia.org/pipermail/wiki-research-l/2008-November/000697.html

---

21
Figure 4: The program *history flow* can visualize the evolution of an article and easily give an overview when major changes occur: The visualization shows an anonymous user making a major extension of the Danish Wikipedia article about Bobby Fischer on April 20th 2005. However, ‘Zorrobot’ makes the most changes. This Norwegian program automatically updates links to Bobby Fischer articles in other language versions of Wikipedia.

The Minnesota study had found that by July 2006 more than 40% of the persistent word views where made by users in the top 0.1%.\(^1\)\(^2\)

Counting of anonymous, bot and sock puppet editing as well as the name spaces investigated may confound such conclusions: Are anonymous edits infrequent editors or a registered user not logged in? Is a specific IP-number a network proxy merging multiple authors? Are some of the edits actually automatic edits from an unregistered bot?

Bots and edits via assisted editing tools make up a considerable part of the total number of edits. In 2007 Stuart Geiger found that the number of edits of this type is larger than the number of edits made by anonymous editors. On one particular page such edits made up 75% of the total edits,\(^1\)\(^8\),\(^1\)\(^0\) see also page 37.

Even in dedicated teaching wikis where students are given writing assignments the level of contribution may be highly variable.\(^2\)\(^1\),\(^1\)\(^2\)

**User characteristics**

A Pew Research Center’s *Internet & American Life Project* survey among adult Americans based on telephone interviews found close to 80% were Internet users. Among these Internet users education level, age and home Internet connection type were the strongest predictors for Wikipedia use. Household income and race/ethnicity were also predictors for Wikipedia use, while gender was not a strong predictor: 56% of male and 50% females Internet users used Wikipedia.\(^4\)

In April 2011 the Wikimedia Foundation surveyed Wikipedia editors with the results published in a 75-page report and they reported several user characteristics for the over 5,000 respondents:\(^2\)\(^1\)\(^3\) One question is whether one can claim that Wikipedia is built by mostly non-experts or in the words of Andrew Lih ‘a bunch of nobodies’. For the highest level of education 8% survey respondents reported a PhD while 18% reported a Master. A large portion of the editors feel proficient with computers with 36% reporting that they program &
create their own applications. In an interview study among 32 contributors to health-related Wikipedia articles 15 (47%) worked in a health-related fields and mainly as clinicians.\textsuperscript{214}

Wikipedia contributors have a sizable gender gap. Quite a number of surveys have addressed the issue: A 2015 blog post listed 12 surveys.\textsuperscript{215} Among the surveys is for instance the Wikipedia Survey published in 2010. It found that only 13\% of contributors were female.\textsuperscript{216} Another survey, the Wikimedia Foundation editor survey of 2011, found 9\% of the respondents to be females.\textsuperscript{213} The MediaWiki software allows users with an account to set their gender in the preferences. Robin Peperman’s Toolserver application ‘User preference statistics’ aggregated the statistics across users, e.g., on the Danish Wikipedia it showed that in February 2012 3'153 users declared themselves male and 677 female, i.e. 82.32\% and 15.68\%, respectively. Lam et al. used the user preference setting statistics on the English Wikipedia and also userboxes on user pages specifying gender. They found ratios in a similar range: Around 13\% for the userbox method and around 16\% for the preference setting method.\textsuperscript{217} Furthermore, they found that the ratio has remained more or less constant in the time window they looked at (2006–2011). The reason surveys have reported somewhat lower ratios than the user preference setting statistics may be due to survey non-response bias, and Hill and Shaw argue that the survey ratios should be around 25\% higher.\textsuperscript{218}

One sees the approximately 85-to-15 percent ratio in public thought-leadership forums as monitored by The OpEd Project.\textsuperscript{219} However, the community on free and open source software (FOSS, sometimes Free/Libre and Open Source Software, FLOSS) has a much larger gender gap: a survey found only 1.1\% females in the sample.\textsuperscript{220} Various reasons for the gender gap can be put forth, e.g., that the Wikipedia editing is somewhat technical and that women are ‘scared away’ because of hostile discussions in the online community of Wikipedians. Indeed the Wikimedia Foundation editor survey of 2011 found that, e.g., 4\% of the responding female editors reported ‘I was stalked online’.\textsuperscript{213}

For the WikiWomenCamp May 2012 Laura Hale and others summarized the involvement of women in the different Wikimedia projects across countries, see WikiWomenCamp/FAQ/Perspectives article on the Meta-wiki.

A 2007 study provided evidence that young patients with schizotypal personality disorder would use the Internet for social interaction significant more than controls.\textsuperscript{221} Does Wikipedians have a special personality type? An Israeli study sought to answer this question subjecting 139 Wikipedians and non-Wikipedia to a personality questionnaire.\textsuperscript{222} In the study Wikipedians scored lower on ‘Agreeableness’ and higher on ‘Openness’. Scores on ‘Extroversion’ and ‘Conscientiousness’ personality dimensions depended on the sex of the subject. Based on previous research they hypothesized that Wikipedians would score lower on extroversion, but their results indicated that only female Wikipedians would score lower. The authors wondered why Wikipedians scored lower on agreeableness, suggesting that contribution to Wikipedia as an apparent prosocial behavior, links to egocentric motives such as “personal expression, raising self-confidence, and group identification”.\textsuperscript{222}

One of the major dichotomies of user characteristics with respect to Wikipedia editing behavior is concerning inclusionists and exclusionists/deletionist, — inclusionists being contributors believing “that Wikipedia should contain pretty much anything, as long as it’s factual and verifiable”,\textsuperscript{25} while exclusionists being more stringent on what go into the encyclopedia.

With the use of infoboxes or explicit categorization on their user page users may manifest that they belong in one or more user categories. There is a sizable network of user categories, for the English Wikipedia, see the subcategories from the top at Category:Wikipedians. Some users explicitly declare themselves as, e.g., inclusionists or Hindu Wikipedians. The www.wikichecker.com website made analysis results available of some of the user categories across a number of language versions of Wikipedia. Among the results one finds that of 5'973 users manifesting their beliefs in total on the English Wikipedia 7\% declare themselves to be Jewish and 38.9\% to be Atheists (see Figure 5), while 20.1\% of 4,433 manifesting their editing phi-

\textsuperscript{xxii}http://www.wikichecker.com/religion//, February 2014.
losophy declare themselves to be inclusionists.xxiii

Geography

An IP number can be resolved to a geographical location (“geo lookup”). Wikimedia data analyst Erik Zachte used information from server logs, converted it to a geographical coordinates and constructed a dynamic visualization of the edit pattern through time with world map background.xxiv There are websites with similar dynamic visualization. László Kozma’s WikipediaVision runs from http://www.lkozma.net/wpv/. Wikipedia Recent Changes Map was built by Stephen LaPorte and Mahmoud Hashemi and available from http://rcmap.hatnote.com in May 2013. It captured edits from unregistered users on different language versions of Wikipedia in real-time and displayed them on a map. While these Web services focused on recent changes edits Federico Scrinzi, Paolo Massa and Maurizio Napolitano’s Wiki Trip website would display geo-resolved unregistered edits based on data from single Wikipedia articles through time with an OpenStreetMap map as background.

With access to page view data, Erik Zachte constructed the WiViVi service with a map of page view with respect to Wikipedia language version.xxv Students from the Technical University of Denmark constructed a Web service that would download the blocked IP addresses information from Wikipedia (or any other MediaWiki installation) and after a geo lookup render the geographical distribution of the blocked IPs with a temporally static heat map visualization.

By examining the user page with natural language processing techniques researchers determined the country associated with a user. They further examined geotagged English Wikipedia articles associated with the Middle East and North Africa (MENA) region and found that Western-based editors formed the majority of contributors to the MENA articles.xxvi Another investigation of the geotagged articles showed that across 44 language versions of Wikipedia more than half of the 3336 473 geotagged articles had a geotag inside a small circle encompassing all of western Europe, southern Scandinavia and parts of Eastern Europe, i.e., showing a geographically uneven coverage of Wikipedia.xxvii

Organization and social aspects

How is Wikipedia organized? In a critical news comment from 2008 Seth Finkelstein characterized it as “a poorly-run bureaucracy with the group dynamics of a cult”, an “oligarchy” and having “an elaborate hierarchical structure which is infested with cliques and factional conflicts”.xxviii Researchers describe Wikipedia and its relation to concepts such as “community”, “social movement”, “benevolent dictator” and “network sociability”.xxix, xxxi, xxxii, xxxiii Several studies refer to a “core group of dedicated volunteers” or “critical mass”, and Wikipedia has been regarded as governed by a so-called “benevolent dictator” or “constitutional monarch” (Jimmy Wales).xxxi, xxxii, xxxiii, xxxiv The new information and communication technologies have been regarded as creating “network sociality” rather than a “community”, — not based on a common narrative and eroding enduring relationships.xxxv

Papers compare the Wikipedia development model with FOSS development.xxxvi, xxxvii In his classic text on FOSS development, The Cathedral and the Bazaar,xxxviii Eric Raymond came up with the concept of bazaar-type development as opposed to cathedral-type development, with bazaar development characterized by an open development process making early and frequent releases of the developed code. Wagner has compared Wikipedia work with the bazaar approach and argues that Wikipedia fit many of the features of bazaar-style development.xxxix On the other hand Magrassi argues that FOSS requires more top-down coordination than Wikipedia because that software needs to be coherent while Wikipedia can have low coherence: A Wikipedia article may still be good even if other articles are of low quality or non-existent. In major FOSS systems one may see a hierarchical structure with a “benevolent dictator” or “core developers” on the top, followed by “co-developers” and ordinary developers, active users and passive users.xxxix, xxxv, xxxvi For the “development” of an article on Wikipedia this is not necessary. Wikipedia may meet face-to-face and Wikipedians can coordinate work in WikiProjects. However, most of the construction process of Wikipedia takes the form of an indirect collaboration, where individual members in a group make small changes to a shared structure inspiring other to improve it even further, — to paraphrase Peter Miller.xxxi Indeed researchers have frequently in-

xxiv The visualization was available from http://stats.wikimedia.org/wikimedia/animations/requests/- AnimationEditsOneDayWp.html and a blog post had background information http://infodisiac.com/blog/2011/05/wikipedia-edits- visualized/
xxv https://stats.wikimedia.org/wikimedia/ animations/pagerview/wivivi.html
voked Pierre-Paul Grassé’s concept of *stigmergy*—originating in the study of animal behavior—and applied it on the Wikipedia process. Mark Elliott would argue that “collaboration in small groups (roughly 2–25) relies upon social negotiation to evolve and guide its process and creative output” while “collaboration in large groups (roughly 25–n) is dependent upon stigmergy” admitting that in the stigmergic collaborative context of Wikipedia social negotiation can take place (e.g., article discussion, email, Internet relay chat), but take a secondary role. He also sees stigmergic wiki collaboration as distinct from “co-authoring”, where his idea of “co-authoring” consists of social negotiation in the creative gestation period.

Wikipedia has seen an increasing institutionalisation with Wikimedia Foundation, Arbitration Committee and Association of Members’ Advocates.

The Wikimedia Foundation has the primary role of running the computer that serves Wikipedia and its sister projects. Other organizations may also affect Wikipedia. The transition from GFDL to CC licence involved the Wikimedia Foundation, the Free Software Foundation (FSF), Wikimedian voting and Wikimedian Foundation Board of Trustees decision. FSF also acts as a license steward.

Other issues that the research literature has discussed are privacy and legal aspects (copyleft license).

### Popularity

A number of companies, such as comScore and Alexa, collect usage statistics from Web users, and such statistics allows Wikipedia to be compared to other Web sites. The September 2006 traffic got Wikipedia Sites listed as number sixth on comScore’s worldwide ranking, and in October it broke into U.S. comScore top ten. Alexa put wikipedia.org as number 7 in traffic rank in September 2009. Stats.wikipedia.org/reportcard reports the temporal statistics from comScore statistics, and it, e.g., shows that throughout most of the first half of 2010 Wikipedia had 350 million unique visitors and around 8 milliard page requests. In January 2016, Wikimedia Foundation chose to stop using the comScore statistics as they believed the data were “no longer fully representative” of their traffic.

Pew Research Center’s Internet & American Life Project has conducted surveys on Wikipedia use in 2007 and 2010. In May 2010 42% of adult Americans used Wikipedia. This was up from 25% in February 2007. At 24% Wikipedia is by far the most used educational and reference site. In 2007 Yahoo Answers was second (4%), while, e.g., Google Scholar was on 1%. In 2010 use of Wikipedia among Internet users was more popular (with 53%) than instant messaging (47%) and rating a product, service or person (32%), but less popular than using social network sites (61%) or watching online videos (66%).

Another way of capturing statistics related to usage is through the Google Trends Web service that displays search volume for the Google search engine through the years based on a query, see Figure 7. Users use “wikipedia” and “wiki” to refer to Wikipedia, but “wiki” may also be used to refer to other wikis related to, e.g., Minecraft, Naruto and Skyrim. There have been steady declines in the “wikipedia” search volume since the middle of 2010 and for “wiki” since September 2011. The declines are not necessary an indication of a decline of the use of search engine to access information on Wikipedia or other wikis. Given that Wikipedia articles rank high in search engine results, one may speculate that Internet users have come to expect this and not finding it necessary to explicitly mention Wikipedia in the search query. The Google Trends has also be used to collected data related to individual Wikipedia articles. Gene Wiki researchers correlated Google Trends and Wikipedia article views for genes.

An analysis of the Arabic blogosphere identified the English Wikipedia as the second most often linked to site from Arabic blogs and only surpassed by YouTube. After Al Jezeera, BBC and Flickr the Arabic Wikipedia ranked sixth in received citations.

Several researchers have investigated the Internet search ranking of the articles of Wikipedia. Using search engine optimization techniques two researchers investigated the Google ranking of the English Wikipedia for health topics. The queries were 1726 keywords from an index of the American MedlinePlus, 966 keywords from a NHS Direct Online index and 1173 keywords from an American index of rare diseases (U.S. National Organization of Rare Diseases) and compared Wikipedia to .gov domains, MedlinePlus, Medscape, NHS Direct Online and a number of other domains. They found the English Wikipedia as the Web site with most top rankings.

Wikipedia’s Internet search engine ranking on individual queries may depend on the type of query whether it is navigational, transactional or infor-
A 2012 survey found that almost 50% of 173 Danish medical doctors used Wikipedia search for information related to their profession.\textsuperscript{248,249} The survey recruited subjects via the online network DoctorsOnly which might have biased the results.

Detailed analysis of the individual articles of the English Wikipedia shows that bursts of popularity on individual pages occur during Super Bowl halftime entertainment and around the death of a famous subject, e.g., Whitney Houston, Amy Winehouse and Steve Jobs. Other causes of increased view of particular pages are the so-called Google Doodles, denial of service attacks on particular pages, second screen and Slashdot effects.\textsuperscript{250} Page views on individual articles are highly skewed, see Figure 8.\textsuperscript{250,251}
Economy

How is the economics of Wikipedia? What is its economical value? How much would users be willing to pay for it? What would it cost to reconstruct Wikipedia? In 2013 these issues were considered in a short review by Band and Gerafi, that concluded that the economic value of Wikipedia was in the range of tens of milliards and consumer benefits hundreds of milliards,252 see Table 6. One study included in the review estimated the market value by comparing Wikipedia to other Internet companies in terms of reputation and unique visitors. This economic triangulation with LinkedIn, Twitter and Facebook put the market value of Wikipedia to between $10 and $30 milliard (United States Dollars). Other methods estimate the economic value of Wikipedia between $6.6 and $1’080 milliard. The extreme maximum value comes from considering a Wikipedia reader session as a ‘research question’ and using a value of a ‘research question’ on $45 from the National Network of Libraries of Medicines. Given the popularity of entertainment and pop cultural phenomenons on Wikipedia253 and that, e.g., students report using Wikipedia for ‘entertainment or idle reading’254 it seems questionable that the majority of Wikipedia sessions should be regarded as valuable as a 45-dollars research question.

Estimates of replacement costs may use estimate of labor hours. Geiger and Halfaker has put one such estimate forward for data on the English Wikipedia up until April 2012 reporting a value of a bit over 40 million hours. Extrapolating to all Wikipedias they reported a number on approximately 103 million labor hours.255 Another Internet phenomenon may provide comparison: Korean pop artist PSY’s 4:12 minute YouTube video Gangnam Style reached 2 milliard views in May 2014 corresponding to 140 million hours, thus the hours used on building all the Wikipedias compare roughly with time spend on watching the (as of 2014) most popular YouTube video. For the double-digit milliard market value estimate from advertising it is worth remembering that just a small note about the potential of advertisement (“Bomis might well start selling ads on Wikipedia sometime within the next few months, […]”) spawned the Spanish Fork in 2002.25

Edit-a-thons, where (unpaid) editors come together writing Wikipedia articles, may have a variety of goals, such as writing new articles, extending or discussing existing, socializing among Wikipedians, recruiting new editors or increasing general awareness. In 2013 Sarah Stierch from the Wikimedia Foundation surveyed a number of these edit-a-thons.258 Some of the surveyed program leaders tracked budget and donations for the event. Stierch found an average edit-a-thon budget on around $360, and for the 5 events where she had sufficient data she could compute a ‘production cost’ of around $17 for each ‘printed page’ (equivalent of 1’500 characters). English Wikipedia articles had

Table 6: Economic value of Wikipedia based on an 2012 overview by Band and Gerafi.252

<table>
<thead>
<tr>
<th>Milliard USD</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10–30</td>
<td>Market value from comparison with other social media sites.</td>
</tr>
<tr>
<td>21–340</td>
<td>Market value based on income from potential fee paided by reader subscribers times revenue-value ratio</td>
</tr>
<tr>
<td>8.8–86</td>
<td>Market value based on possible advertising revenue</td>
</tr>
<tr>
<td>6.6</td>
<td>Replacement cost based on a $300 per article cost</td>
</tr>
<tr>
<td>10.25</td>
<td>Replacement cost based on estimates of labor-hours and a $50 per hour cost</td>
</tr>
<tr>
<td>0.63/year</td>
<td>Estimate for cost with full-time paid writers employed to update Wikipedia</td>
</tr>
<tr>
<td>16.9–80</td>
<td>Annual consumer value based on received benefit that could have had a fee</td>
</tr>
<tr>
<td>54–1’080</td>
<td>Annual consumer value based on comparison with librarian research question cost</td>
</tr>
</tbody>
</table>

xxviii Programs: Evaluation portal/Library/Edit-a-thons
an average of 3,655 bytes per article in January 2010.\textsuperscript{xxix} Other language versions usually have a lower byte count, so a rough estimate of an average Wikipedia article is between 1 and 2 printed pages. With $25 per Wikipedia article and 30.8 million Wikipedia articles as the official January 2014 count the replacement cost is only $770 million with unpaid ‘edit-a-thon’ writers, — considerably lower than any of the values presented by Band and Gerafi when they computed the estimated replacement cost of all of Wikipedias.\textsuperscript{252, 257}

The Wikimedia Foundation gets donations throughout the year, and it makes some of the donation data available, e.g., from the Web server frdata.wikimedia.org. The donation data has not caught a major share of researchers’ attention, — if any at all.

Why do people edit?

What motivates Wikipedians? Wikipedia itself has a section in the “Wikipedia community” English Wikipedia article discussing the motivation of volunteers giving pointers to research on the issue. Much of the research in this area is based on qualitatively analyzed interviews of Wikipedians.\textsuperscript{20, 258, 259} However, the area has also seen mathematical modeling being employed on edit histories and survey data.\textsuperscript{260, 261} Others have conducted ‘field’ experiments on Wikipedians or relied on events for natural experiments.\textsuperscript{262, 263}

Some Wikipedians mention in interviews that they started editing Wikipedia because they discovered errors or omissions in an article that they knew something about.\textsuperscript{258} Does this means that the more complete and complex Wikipedia becomes the more difficult it will be to attract new editors?

The attraction of Wikipedia is in part due to the low barrier for entry,\textsuperscript{258, 264} also referred to as low opportunity cost.\textsuperscript{19} Introducing required user registration and ‘sighted’ or ‘flagged revisions’ would heighten the barrier and possibly result in less recruitment of new users. The wiki markup could form a barrier for entry, and Wikimedians have hoped that a WYSIWYG editor would increase take-up of new contributors, and indeed a rational for developing of the of the so-called VisualEditor stated: “The decline in new contributor growth is the single most serious challenge facing the Wikipedia movement. Removing the avoidable technical impediments associated with Wikimedia’s editing interface is a necessary pre-condition for increasing the number of Wikipedian contributor”.\textsuperscript{265}

But why do Wikipedia contributors keep on editing? The same question may be asked for free and open source software programmers working for free. Some researchers view Wikipedia as a public good where the contributors “devote substantial amounts of time with no expectation of direct compensation” and a Wikipedian’s participation seems “irrationally altruistic”.\textsuperscript{225}

Self-interest may lay behind some contribution: A software program can solve a specific problem for its creator, and a Wikipedia entry can organize information relevant for the editor to remember. But Wikipedia editors not only work on information directly relevant for themselves. Surveys among Free and Open Source developers mention skill development as the most important factor for beginning and continuing.\textsuperscript{37} Once the skills are developed the software showcases the skills for potential employers, so the contribution can be seen as a form of self-marketing on the job market.\textsuperscript{225} For Wikinews, Jimmy Wales noted that contributors recognized it “as a site, where they could learn to become a journalist”.\textsuperscript{124}

It is unclear if Wikipedia contribution has the same skill signaling value as free software development. Andrew George writes that the potential is “simply nonexistent”.\textsuperscript{225} Nonexistent is not entirely true: A few Wikipedians have gained temporary employment as Wikipedian in Residence. Such positions typically require thorough experience with Wikipedia. It is possible that experienced Wikipedians may also look forward to a job with paid editing. The controversial company Wiki-PR boasts of having a “network of established Wikipedia editors and admins” available to edit Wikipedia on behalf of companies.\textsuperscript{266} A few people are professionally employed to work on Wikipedia. On the Danish Wikipedia articles on runic stones have been made by professionals as part of a job. In her survey among 1’284 public relations/communications professionals DiStaso found that 31% of the respondents had edited in their client or their company Wikipedia page.\textsuperscript{260} It may be worth to note that now company paid contributions form the majority of work carried out on the FOSS Linux kernel: Only between 13.3% and 16.3% of Linux kernel changes come from contributors without corporate affiliation.\textsuperscript{267} thus there is a marked difference between FOSS and Wikipedia labor market.\textsuperscript{268}

After interviewing 22 volunteer Wikipedians Forte and Bruckman likened the incentive system of Wikipedia to that of the scientific community,
where the “cycle of credit” is the most important aspect in the incentive system and prestige is earned over long periods of time. Attribution of authorship may on the surface seem to be a difference between the science world and Wikipedia (authors of Wikipedia articles are not displayed prominently on the page). However, Forte and Bruckman find that through the Wikipedia article editing history contributors recognize one another and that Wikipedians also often claim “ownership” of articles. Another comparison of Wikipedia and academic writing was much more critical: Seth Finkelstein commented on Wikipedia prestige and strongly criticized Wikipedia claiming that “it fundamentally runs by an extremely deceptive sort of social promise.”

Yang and Li attempted to contact a random sample of 2'000 users on the English Wikipedia. Receiving 219 valid responses to their questionnaire they used structural equation modeling to describe knowledge sharing behavior in terms of intrinsic motivation, extrinsic motivation, external self-concept and internal self-concept. They found that internal self-concept-based motivation was the most important factor for the knowledge sharing behavior, and this factor was associated with questions such as “I consider myself a self-motivated person” and “I like to share knowledge which gives me a sense of personal achievement.” On the other hand intrinsic motivation was found to rarely motivate. This factor was associated with questions such as “I enjoy sharing my knowledge with others” and “Sharing my knowledge with others gives me pleasure.” Other researchers have also distinguished between intrinsic and extrinsic motivations. Müller-Seitz and Reger mentions low opportunity costs, reputation among peers, future career benefits, learning and development and contributions from the community as extrinsic motivations while creative pleasure, altruism, sense of belonging to the community and anarchic ideas as intrinsic.

Andrew George proposes the state of ‘flow’ as an intrinsic motivation for Wikipedians, altruism, community identification and peer recognition. Based on interviews Rosenzweig noted that many Wikipedians had a passion for self-education.

Barnstar awards are a form of peer recognition in Wikipedia, and a few studies have analyzed this element with respect to motivation. In the experiment of Restivo and van de Rijt reported in 2012 the researchers split 200 productive Wikipedians in two groups and awarded a barnstar to the
Figure 9: Effect of barnstar awards on productivity of Wikipedians: Figure 1 from Restivo and van de Rijt (2012). “Experiment” is median productivity of 100 barnstar awarded Wikipedians, while “control” is matched Wikipedians that received no awards from the experimentors. Third-party awards are displayed with triangles. © Restivo and van de Rijt. CC-BY.

Wikipedians in one of the groups (“experiment” in Figure 9) and none in the other group (“control”). The productivity decreased in both groups, but the productivity was sustained better in the group with awarded Wikipedians, see Figure 9. Benjamin Mako Hill and Aaron Shaw would also see an effect of barnstars summarizing their findings with: “We show barnstars can encourage editors to edit more and for longer periods of time, but that this effect depends on the degree to which the recipient chooses to use the award as a way to show off to other editors”.

In his classic work on FOSS development, *The Cathedral and the Bazaar*, Eric Raymond also discusses motivation for voluntary work and mentions the concept of “egoboo”, a shortform of ego-boosting. He argues that science fiction fandom “has long explicitly recognized ‘egoboo’ (ego-boosting, or the enhancement of one’s reputation among other fans) as the basic drive behind volunteer activity.” The existence of public edit counter tools as well as pages such as Wikipedia:List of Wikipedians by number of edits allow Wikipedians to ‘egoboost’ via their ‘edit count’.

The notion of addiction in Wikipedia contribution (colloquially ‘wikipediaholic’, ‘wikiholic’ or ‘wikipediolution’) has also been noted. New York-based Wikipedia researcher Dan Cosley even claimed “It’s clearly like crack for some people.” Finnish researchers proposed gateway theory, originating in the study of drug use, as a framework for explaining online participation in FLOSS development and Wikipedia contribution (cautioning that they do not claim that “online participation, at large, would qualify as addiction”).

Contrasting with antecedent-based explanations, the gateway theory-based argument notes that the initial online participation occurs ‘due to chance’ without prior planning.

Evolutionary biological explanations of religion may involve the idea of religious rituals as reliable costly signals between adherents for social commitment. Should the edit count partially be explained as a hard to fake, costly signal between the adherents of the ‘Wikipediaistic’ religion?

**Why do people leave?**

Studies of why editors leave Wikipedia are much less performed than studies on why editors start out. In 2009 Wikimedia Foundation staff and volunteers developed an online survey and emailed it to 10'000 contributors. The included contributors should have had between 20–99 edits and no edits in the past three months. The survey conducted in the beginning of 2010 had a response rate on 12% from 1238 editors. Around half left due to personal reasons (e.g., other commitments), another half left due to “bad terms”. The researchers split the “bad terms” into reasons related to complexity (e.g., writing a encyclopedia is difficult) and community (e.g., other editors being rude). Among the high scoring statements for leaving were “my work kept being undone” and “several editors were too stubborn and/or difficult to work with”.

Eric Goldman speculates that Wikipedia has a particular vulnerability life changes among its contributors, because they tend to be young, unmarried and childless. The Wikimedia Foundation editor survey of April 2011 found ‘less time’ with 37% as the most reported reason for becoming less active on Wikipedia, compared with, e.g., 7% for ‘conflict with other editors’.

Based on large-scale quantitative analysis of, e.g., the editing pattern and its temporal evolution the researchers concluded that Wikipedia had a slowing growth up to 2009. They explained this phenomenon with “limited opportunities in making novel contributions” (the low-hanging fruits are already taken) and “increased patterns of conflict and dominance due to the consequences of the increasingly limited opportunities”.

An increasing ‘xenophobia’, where seasoned Wikipedia editors under constant spam and vandalism threat develop a ‘revert first’ mentality and away from ‘assume good faith’, may make it harder for contribution from unsophisticated users to stick.
aspect affects the retention of new contributors.

Contributor may also leave because of project forking, examples being Wikitravel and the so-called Spanish Fork of Wikipedia.

**Edit patterns**

Several studies analyze and models the edit patterns of users and articles.

One claim for the number of new edits to an article within a time interval comes from Wilkinson and Huberman, who hypothesized that the number of new edits in the time interval $\Delta n(t)$ is on average proportional to the total number of previous edits with some amount of fluctuation

$$\Delta n(t) = [a + \xi(t)] n(t),$$

where $a$ is a constant representing the average rate of edits and $\xi(t)$ is the fluctuation. Via the central limit theorem they arrive at the lognormal distribution for the number of edits

$$P[n(t)] = \frac{1}{n\sqrt{2\pi s^2t}} \exp \left[ -\frac{(\log n - at)^2}{2(s^2t)} \right],$$

where $\mu = at$ and $\sigma^2 = s^2t$, i.e., the mean and variation of this lognormal distribution it linearly related to the age $t$ of the article. They show on a robot-cleaned dataset from the English Wikipedia that the distribution of the number of edits follows the model well except for a few events, e.g., one is the edits of US town articles. It is fairly surprising that the number of edits can be modeled with such a simple model that only contains Wikipedia-intrinsic variables. No external forcing variables, e.g., number of editors, ‘importance’ or ‘notability’ of article topic, appear in the equation. Note that the data set only covers the period where Wikipedia had exponential growth and making it how well the model fits newer data.

The overall temporal evolution of the number of articles and editors are perhaps the most important statistics of Wikipedia. It is regularly computed by Wikimedia Foundation data analyst Erik Zachte and reported on the Wikimedia Statistics homepage [http://stats.wikimedia.org/](http://stats.wikimedia.org/). Of constant worry for Wikimedia is the statistics of the number of active editors and the editor retention. The decaying number of active editors on the English Wikipedia has been apparent for a number of years and given rise to the so-called ‘oh shit’ graph, see Figure 10, which shows the number of active editors to peak in 2007 and then decaying. The number has continued to decrease and was in the beginning of 2014 around 30’000 according to the Erik Zachte data analysis.

Wikipedians have reported mathematical models of growth of the number of articles on the English Wikipedia and made them available from [Wikipedia:Modelling Wikipedia's growth](https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia%27s_growth). Between 2003 and 2007 an exponential model fitted the data well. A logistic model estimated from data available in 2008 indicated that Wikipedia would reach a maximum around 3 to 3.5 million articles, a result reported by Suh et al. in 2009. As the logistic model does not model the data well at least after 2010 other models have been suggested and a seasonal variable have been included in the model. A model involving the Gompertz function, $y(t) = ae^{bc^t}$, was reported to project a maximum of about 4.4 million article, — a number which the English Wikipedia has surpassed. Suh et al. suggested a hypothetical Lotka-Volterra population growth where the growth was not bounded by a constant but limited by a linear function of time arguing “there is a general sense that the stock of knowledge in the world is also growing.”

**Why does it work?**

Why does Wikipedia not succumb to vandalism and opinionated content? There are other open fora on the Internet: The Usenet and emails. The Usenet is like Wikipedia a public space, but from a start among ‘geeks’ it has been used for pornography and pirated content. Another collaborative knowledge building project, [Open Directory Project](https://opencyc.org/), also relied on free labor, but was in 2006 declared to be ‘effectively worthless’ by Eric Goldman.

In 2011 Benjamin Mako Hill presented his research about Wikipedia and other crowd-sourced...
online encyclopedia, Interpedia, The Distributed Encyclopedia, Everything 2; h2g2, Aaron Swatch’s The Info Network, Nupedia and GNUPedia, sought to answer what distinguished the successful Wikipedia from the failed or less successful projects. Hill noted that Wikipedia offered low transaction cost in participation and initial focus on substantive content rather than technology. Two of the earliest comment on the Wikipedia phenomenon pointed to the low transaction cost as a reason for the Wikipedia success. Philippe Aigrain noted low transaction costs as an essential factor for success, — not only for Wikipedia but for other “open information communities”, e.g., Slashdot. Aigrain would see transaction costs as having a number of aspects apart from (possible) monetary costs: cognitive, information, privacy, uncertainty and locking-in costs, with, e.g., the cost of “navigating the transaction management layers” as an information cost. Aigrain also argued for two other factors contributing to the Wikipedia success: a clear vision and mechanisms to fight hostile contributions. He sees the statement on neutrality of point of view as a clear vision and the revision control system as the key mechanism to counter hostile or noisy contributions, and indeed it seems very difficult to imaging how Wikipedia could work without its revision control system where contributors can restore ‘blanked’ articles. There are likely a large number of important technical elements in the MediaWiki software that lower transaction costs for ‘good’ contributors and heightened it for ‘bad’: CAPTCHAs, ‘autoconfirmed’ user level, user blocking, IP range blocking, blacklists, new pages patrolling and watchlists. The flagged revisions of the German Wikipedia might be the reason for its high quality. The ability of Wikipeadians to communicate beyond the stigmergic article space collaboration, e.g., via discussion pages, is presumably important, — at least they are almost always used for high quality ‘featured articles’.

As a newder development of a Wikipedia challenger, Google launched the Knol service in 2008, and while contributors could monetize their content Google announced the closing of the service in 2011. Although a quantitative study could find differences in the content of articles of Google Knol and Wikipedia, e.g., with respect to the number of references, it seems unclear if we can blame such differences for Google Knol’s demise. One commentor noted that “Google didn’t seem to allocate many resources to the project, and it certainly didn’t put much emphasis behind it from a press standpoint”. Before Knol and with likely inspiration from Wikipedia Microsoft unsuccessfully opened for user contribution to their encyclopedia Encarta. Andrew Lih noted “it was not timely, open, social, or free.” and a “half-hearted implementation”.

I have not found mentioned that Wikipedia is an excellent example of hypertext. Much quality text on the Web is not linked very much, e.g., a typical New York Times news article has typically no links in the body text. Web scientific articles rarely has links. Google Knol had typically no hypertext links besides table of content-type or external references. Even MediaWiki-based Scholarpedia has few intrawiki links. URLs in Danish encyclopedia Den Store Danske and Google Knol are/were not well-predictable. Links on the ordinary Web based on non-wiki technology do not display the availability of the linked page. The URLs of Wikipedia articles are well-predictable, likely reducing navigational costs.

The creation of Wikipedia is a decentral process without economic incentives, where individual contributions get aggregated to a collective result. Certain Web-games have the same character: In the Web-based market game Hollywood Stock Exchange the goal is to predict the box office of a movie by buying and selling virtual shares. Four weeks after the opening the Hollywood shares get the toy money in accordance with the earnings of the movie. The price of the stock then becomes a prediction on the real earnings of the movie. Artificial market games perform surprisingly well in predictions, e.g., Hollywood Stock Exchange could predict Oscar winners better than individual experts. Apparently the prediction markets attracts well-informed and well-motivated players whose mutual trade creates information, even though the participants does not gain any money. There is not real trade in Wikipedia, but Wikipedia forms a decentral knowledge aggregation, like the knowledge markets.

The ‘power’ of the collective has often been invoked and sometimes under the name Linus’ Law, that Eric Raymond coined in The Cathedral and the Bazaar to characterize an element in FOSS development: “Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix obvious to someone,” i.e., openness gives robust results. The concept has also been applied on Wikipedia to explain its success. Greenstein translates the ‘law’ to the Wikipedia context with: “Enough attention to such details [facts and commonly accepted knowledge presented on Wikipedia] elicits the relevant objective information and many viewers then correct errors as needed.” and “[given enough cheap storage, all viewpoints can be represented.” He also notes its difficulties with costly
verification and subjective information where Linus’ Law might not work.

3 Using Wikipedia

Users meet not only Wikipedia as readers of the Wikipedia site. Other web-sites utilize Wikipedia content to enhance their own site, and the rich structured annotations with links, tags and templates allow researchers and developers to use Wikipedia as a knowledge base. Especially researchers in information retrieval and automatic text processing benefit from Wikipedia. In this kind of research WordNet has stood as the foremost open computer-based lexical database for the English Language. Some research compares or combines WordNet and Wikipedia. The review Mining meaning from Wikipedia from 2009 gives an overview of studies using Wikipedia data in natural language processing, information retrieval and extraction as well as ontology building.\(^9\)

Other wikis besides Wikipedia provide data for this kind of research. Wiktionary, the dictionary-like sister project, supplements Wikipedia, and several studies have used information from Wiktionary.\(^295\text{--}298\)

Information retrieval

A program uses Wikipedia articles as references to detect so-called “content holes”. Content holes are neglected or missed subtopics in a text. The researchers regarded user generated content from discussion threads and compared that to the sections found in Wikipedia articles by extracted keywords such as proper nouns and numbers.\(^299\text{--}301\)

Web search queries occurring with high frequency often directly relates to Wikipedia pages. Within a large sample of web queries 38% matched exactly a title of an article in Wikipedia.\(^302\) The content and context of the matched Wikipedia page can then be used to automatically expand the query, e.g., Wikipedia pages can form an intermediate between a user query and a collection of books being searched.\(^302\) Researchers have used the redirects in Wikipedia to form a thesaurus for use in document retrieval systems with automated query expansion.\(^303, 304\)

Researchers also use images from Wikipedia in information retrieval systems research.\(^305, 306\) Researchers used part of the French Wikipedia for the demonstration of a combined text and image browsing system with similarity computation.\(^305\)

Thesaurus construction

The comprehensive Rouget’s International Thesaurus\(^307\) lists both the word agency as well as Federal Bureau of Investigation. None of these two articles link to the other directly. Several groups attempt to make systems that will automatically determine the relatedness of such words and phrases based on mining data in Wikipedia or other sources, e.g., WordNet: “synonym search” or “semantic relatedness”.\(^140, 298, 303, 308\text{--}310\)

Chesley et al. used Wiktionary to expand a labeled list of adjectives.\(^296\) For text sentiment analysis they needed the polarity and objectivity of adjectives. They already had a list with a limited number of words manually labeled in this respect. By lookup in Wiktionary they could assign polarity and objectivity to yet unlabeled words by examining the definition of the adjective. The approach also handled words with negated definition.\(^296\) Zesch and Gurevych used the Wikipedia article and category graph in their semantic relatedness method.\(^140\)

Named entity recognition

One of the basic tasks of information extraction, named entity recognition (NER), deals with identifying named entities such as personal names, names of organizations or genes from freeform text. NER often relies on a machine learning algorithm and an annotated dictionary (gazetteer). Several researchers have used Wikipedia for NER or more general keyword/keyphrase/concept extraction.\(^304, 310\text{--}315\)

French Exalead’s Wikifier is a Web service that will take a document (either as a URL or a text) and serve it back as an HTML page and link terms within the document to relevant Wikipedia articles. With named entity recognition it detects people, places or organizations. CC-BY-NC-SA-licensed Accurate Online Disambiguation of Named Entities (AIDA) software for online named entity extraction with disambiguation uses Wikipedia indirectly through YAGO2.\(^315\) A online demo version is available.

A game development system, that starts with a newspaper article, identifies country and notable person based on information in Wikipedia. The system furthermore gauges the sentiment towards the person by using posts from Twitter in sentiment analysis.\(^316\)
Translation

Wikipedia can act as a resource in machine translation and multilingual systems by using the language link that connects different language versions of Wikipedia. Before the Wikidata project interlanguage links was part of the article page of the MediaWiki site through a special markup that usually appeared at the bottom of the page. Daniel Kinzler’s WikiWord system extracts lexical and semantic information from Wikipedia to form a multilingual thesaurus. In an extension to the system called WikiPics he use it in combination with images from Wikimedia Commons for multilingual image retrieval.

It seems obvious that the multilingual Wikidata will become a valuable and easier accessible resource for future translation studies relying of Wikimedia data. Researchers will no longer need to crawl multiple Wikipedias to establish naming variations (through redirects) and extract interwiki language links. Instead all basic language links is readily available in Wikidata in machine readable format. Arun Ganesh (‘Planemad’), one of first to explore Wikidata-based translation, built a prototype with a map of India with labels for the states of India retrieved from Wikidata: The Wiki Atlas. It would allow the user to switch between, e.g., Hindi, English and Zhongwen representations of state names. In 2014 Denny Vrandečić and Google would take this approach a step further by publishing the qLabel Javascript library, enabling a website developer an easy method to provide Wikidata-based multilingual website content.

The Danish company GrammarSoft provide online translation of Wikipedias from one language to another. It has translated the Swedish Wikipedia to Danish and made it available from http://dan.wikitrans.net/ complete with hyperlinks, while an Esperanto-English version is available from http://wikitrans.net/. These system uses Constraint Grammar extended to handle the ‘heavily layotted’ text of Wikipedia. Even though one language version of a Wikipedia articles may be a translation of a Wikipedia article in another language, the wikitext is not readily sentence-aligned. Thereby a potentially valuable resource for training statistical machine translation systems is lost. Parts of the documentation for MediaWiki and some other Wikimedia resources are sentence aligned.

Ontology construction and categories

Social websites may allow users to put free-form tags on content. The relationship between the combined set of tags has been term *folksonomy*. Users tag Wikipedia articles with categories and the users can organize the categories in a hierarchy (actually a directed acyclic graph, usually, but not necessarily a hierarchy). While editors may use free-form categories, the categories that survive are the ones which have support in consensus. Jakob Voß distinguishes between collaborative tagging (e.g., used the del.icio.us website), classification (e.g., Dewey Decimal Classification) and thesaurus indexing (e.g., Medical Subject Headings, MeSH, where records/categories may have multiple broader terms), and he regards the Wikipedia category system as a collaborative thesaurus. The categories together, with categorized articles, have been used for ontology learning in a semi-automated system.

Another way of using Wikipedia begins with an already established knowledgebase and then attempts to match and extract data from Wikipedia for further population of the knowledgebase. Friedlin and McDonald would start with records in the medical terminology database LOINC and with a semi-automated program attempt to find corresponding articles in Wikipedia, so the introductory Wikipedia text could be used for a helping description text for the LOINC records. Reagle and Rhue have described an application of a personal name matching method between multiple biographical works—including Wikipedia—as well as gender determination method in a study of gender bias.

Two Technical University of Denmark students, Magnús Sigurðsson and Søren Christian Halling, used the part of Wikipedia dealing with musical groups for the Internet-based search engine MuZeeker and could group search results according to Wikipedia categories. A further related application ran on Nokia mobil telephones.

Studies on Wikipedia also uses the Wikipedia category network as a form a ground truth to assign top categories to Wikipedia articles. Although the category network is directed the researcher consider the undirected network to avoid disconnected categories, though weighting the graph according to right/wrong direction help categorization.

Databasing the structured content

Several groups have extracted information from the categories and templates of Wikipedia and built...
databases. The YAGO system extracted data from Wikipedia and combined it with WordNet. Their aim is ontology building with facts such as “Elvis” “is a” “rock singer” and “Elvis” “has the nationality” “American”. They note that the semi-structured data of Wikipedia is not directly useful, e.g., the category tree has “Grammy Award” as a subcategory of “Elvis” as the rock singer had won that award, but it would be wrong to infer “Elvis” “is a” “Grammy Award”. This is in contrast to WordNet which has a cleaner graph. When reported in 2008 YAGO had 1.7 million entities and 15 million facts. The number of entities may be compared to the number of terms in Gene Ontology that I in 2013 counted to around 40’000.

The largest effort is probably DBpedia with extraction of Wikipedia templates and the information made available at http://dbpedia.org/ BBC uses among others DBpedia for linking documents across their web site.

With the advent of Wikidata, direct extraction of data from the categories and infoboxes of Wikipedia may still be relevant, but it seems that Wikidata will be a much stronger base.

Trend spotting and prediction

Just prior to the announcement of the choice of the United States vice presidential candidat Sarah Palin’s Wikipedia article saw a high editing activity. The activity was higher than for other candidates, thus giving an indication for the choice of Palin. In another case Paula Broadwell’s Wikipedia page contained the short statement “Petraeus is reportedly one of her many conquests” for several months before the secret relationship became public knowledge and led to an FBI investigation and the resignation of the CIA director. Cyber intelligence systems can use this kind of information for predicting events prior to it becoming public knowledge.

Public Web tools monitor the editing pattern and presents analyzed results based on queries: Craig Wood’s Wikirage Web-site ranks Wikipedia articles based on edit activity measured not only as the total number of edits, but also, e.g., the number of undos and unique authors. This tool is useful to identify edit conflicts which may generates lots of edits. WikiChecker also generates statistics over users and individual articles as well as list highly edited pages. A similar Web-site by Dennis Yurichev shows articles created within the last month, last week and last 24 hours ranked by the number of contributing editors. Such a Web service puts emphasis on trending topics that have not before been in the spotlight. One example is Atifete Jahjaga rising from relative obscurity to President of Kosovo. Thomas Steiner and his coworkers present yet another web service for realtime trend spotting: Wikipedia Live Monitor monitors Wikipedia edits across 42 different language version and detect concurrent edit spikes to report breaking news, e.g., the 2013 Russian meteor event. Steiner has presented a related web service, the Wikipedia Natural Disaster Monitor, focusing on the natural disaster articles on Wikipedia for crisis response.

Sérgio Nunes’ WikiChanges creates an on-the-fly graph of the edit activity through time of one or two Wikipedia articles. One example pointed to compares ‘Barack Obama’ and ‘Hillary Rodham Clinton’ articles and shows most edits for the winner of the United States primary election. Inspired by Maximilian Laumeister’s Listen to Bitcoin, Stephen LaPorte and Mahmoud Hashemi built the Web service Listen to Wikipedia which converted Wikipedia recent change feed to an online dynamic visualization with sound. Circles would appear and bells would sound each time an edit was made.

The Wikimedia Foundation did for some time not make statistics on page views available. However, in December 2007 board member Domas Mituzas announced the availability of hourly statistics for each page from [http://dammit.lt/wikistats/](http://dammit.lt/wikistats/). Such data makes it easy to spot trends by simple counting. The user “Henrik” has presented a useful interactive Web service that renders the statistics in monthly histograms from [http://stats.grok.se/](http://stats.grok.se/). Showing topics of public interests — and disinterest — the user “Melancholie” pointed to an example: “Hurricane Gustav” and “Bihar flood”, both natural disasters of 2008 but one in rich USA, the other in an impoverished state of India. The former reached several hundred thousand page views in August and September while the latter only a few thousands.

Based on the viewing statistics Ed Summers made the Web service Wiktrends available in 2012 from [http://inkdroid.org/wiktrends/](http://inkdroid.org/wiktrends/) showing the top 25 most viewed Wikipedia articles for the past two hours. Since 2008 Johan Gunnarsson has run a similar Web service, also called Wiktrends, first from the Toolserver and later on on the Wikimedia Tool Labs servers with up- and down-trends based on day, week, month or year and across many language versions of Wikipedia. Furthermore, Andrew West makes weekly reports of the 5000 most visited pages on the English Wikipedia available on [one of his user subpages on Wikipedia](http://en.wikipedia.org/wiki/Wikistats). Weekly annotated
versions of the Top 25 articles from the West list are made available in the Wikipedia:TOP25 article. West together with user Milowent have also reported on page view spikes, see page 26.

Working from stats.grok.se June and January 2008 statistics and examining health-related topics with probable seasonal effect—such as frostbite, hypothermia, hyperthermia and sunburn—Laurent and Vickers found a clear effect in the page views. On average the ratio was around 3 between June and January page views for these topics. They also analyzed the page view statistics of three articles describing melamin, salmonella and ricin. These examples were associated with official health alerts in 2008, and page view statistics also shows a marked increase correlating with the announcements. In 2014 a company would publish an analysis of English Wikipedia articles with health care information reporting correlation between article views and related prescriptions and unit sales of medication. Another interesting application would examine how well Wikipedia activity could be used to predict movie box office success. Not only the viewing statistics was used as a feature in the prediction, but also the number of edits, the number of users editing and “collaborative vigor”. Keegan did a related small longitudinal study with prediction of 2014 Academy Awards winners based on the number of edits and the number of unique editors in 2014. He predicted 3 out or 5 correctly. Yet another analysis of the utility of Wikipedia page view statistics for prediction of real-life events showed mixed results for the result of political elections in Germany, Iran and United Kingdom.

Like Wikipedia, the Twitter Web service provides an API for returning content in structured format. The Wikipedia services for trend spotting may resemble some of the many Web services for Twitter postings, and programmers have presented systems combining Wikipedia and Twitter. Tom Scott implemented a Twitter bot with Wikiscanner-like functionality: A bot monitoring Wikipedia changes would track Wikipedia edits from United Kingdom Parliament IP-adresses and output changes to a specific Twitter account (@parliamentedits). Inspired from this system Ed Summers implemented a system for the United States Parliament (@congressedits) and released his code enabling the monitoring of several other parliaments, e.g., Sweden with @RiksdagWikiEdit, Denmark with @FTingetWikiEdit and Canada with @gccaedits. Danish Wikipedian Ole Palnatoke Andersen would collect these Wikipedia monitoring Twitter accounts on his list at https://twitter.com/palnatoke/lists/wikiedit with the number of list members reaching 48 in October 2014.

Altmetrics

Altmetrics is short for alternative metrics and is the study and use of measures for scholarly impact based on activity in social media. Mendeley and CitEULike are useful services, but also Twitter, blogs and Wikipedia have been examined as sources for altmetrics, and there have been a few studies on altmetrics with Wikipedia. In a survey among 71 bibliometrics researchers 33.8% stated that “mentions of or links to your work in Wikipedia” had the potential for article or author evaluation. For the statement “article about you on Wikipedia” 26.8% confirmed its potential. My analysis of outbound links to scientific journals from Wikipedia (using the cite journal template) found that the correlation with the de facto standard for scholarly impact measure was reasonable. Bot-generated content can change the number of citations considerable. Whereas typical citations uses the ref tag for in-text citations, the Protein Box Bot would add citations to the end of the article without using the ref tag. When the bot created and amended thousands of articles in 2008 there was a large increase in structured citations using the cite journal template, but without using the ref tag, see Figure 11. Indeed the results indicate that the bot in 2008 was the creator of the majority of the journal citations in the English Wikipedia. American researcher carried out a similar analysis with the 2010 English Wikipedia dump, but instead based the citation extraction on
the PubMed identifier (PMID) or the Digital Object Identifier (DOI). By analyzing the full revision history of the English Wikipedia they identified a trend for journal articles to be cited faster in later years compared to earlier years.336


Analyzing the matrix formed by intrawiki Wikipedia linking with the PageRank and CheiRank methods Toulouse and Hungarian researchers would examine the ranking of the top 100 universities over time and compare it to Shanghai university ranking. They found the PageRank of the top 10 university to increase over time, and, e.g., Harvard and Columbia at the top.148

As researchers themselves can edit pages and insert mentions or links to their own work (as I have done) Wikipedia citations are susceptible to manipulation of the altmetrics measure. Notability criteria, deletionists and general collaboration among contributors may keep a bound on this problem. Also on the downside of using Wikipedia for altmetrics is the poor coverage of scientists on Wikipedia.89

The Altmetrics.com website features Wikipedia as one of altmetrics sources for paper-level scientometrics. It tracks Wikipedia citations to journal articles by date and user.xxxvi

Geotagged content

Some Wikipedia articles have markup with geographical coordinates. These coordinates can be extracted and used, e.g., with rendered maps as, e.g., in Google Earth and Danish Findvej.dk. Wikimedia with its structured content including geotagging can provide a plentora of information for enriching maps, e.g., OpenStreetMap user abahnver-leih combined Wikidata’s entities with geotags and coat of arms images to render coat of arms on map, see http://osm.lyrk.de/wappen/.

The augmented reality platform Wikitude and its smartphone world browser uses location-based Wikipedia content such that Wikipedia article excerpts are overlaid on the camera image. The Wikipedia article shown is selected based on GPS and compass information in the mobile phone. Another augmented reality browser, Lagar, embeds Wikipedia landmarks.

Other uses

WikiTrivia developed by Alex Ksikes and Philipp Lenssen uses the introductory paragraph from Wikipedia for an online quiz. With the title words blacked out from the paragraph the task is to guess the title from the paragraph. The quiz participant chooses a category, such as politics or animals, before the game begins. The online quiz WhoKnows? also uses information from Wikipedia, but indirectly through DBpedia.337

Researchers have used the edit histories of Wikipedia articles in a study of the performance of system for sharing mutable data.338

4 Extending Wikipedia

The original Web protocol of Tim Berners-Lee has always had the ability to both read and write on the Internet. However, in the first decade of the Web the writing capability was little used. Some Web-based applications enabled collaborative workspaces, e.g., the BSCW system, that let users upload files.339 Another effort, WebDAV, would standardize a protocol for distributed authoring on the Web.340 Wikis begin with the work of Ward Cunningham, with many of the early technical ideas described in the book The Wiki Way:29

The ability of users to edit and create new pages, the easy way to make page link between pages, the free structure of wikis, the notion of a sandbox, page locking, camel-case (wikiwords), the basic edit conventions (the wiki markup) and revision control.

Researchers and developers have suggested many technical extensions for MediaWiki, the Wikipedia software: Software tools that automatically create new content or modify existing, e.g., to fight vandalism, tools that help users (readers and editors), e.g., get an overview or provide the content in another form than the usual way, and extensions with additions to the data representation and interfaces to handle structured content.

Wikipedia editing by bots of assisted editing tools has increased over time. When Stuart Geiger examined the fraction of bot and assisted editing in 2009 he found 16.33% of all edits performed by a bot and assisted editing programs used in around 12.16% of all edits. On a particular page used for vandalism fighting, Wikipedia: Administrator intervention against vandalism (AIV), edits by tools are predominant with around 75% of the edits.181,210
Table 8: Automated article classification. The performance column may indicate different types of performance, e.g., accuracy.

### Automated quality tools

There have been several suggestions and implementations of tools for quality assurance in wikis and Wikipedia particularly, and number of tools are in operation. The tools that operate automatically on Wikipedia are usually referred to as bots and may perform mundane tasks, such as between-language Wikipedia version links (between-language Wikipedia links are now taken over by Wikidata) before the advent of Wikidata, add dates to templates, include titles from cited external pages, etc.\(^{344}\) The most important bots are probably the anti-vandalism tools (vandal fighter bots) on Wikipedia. These kinds of bots have been operating since user Tawker’s Squidward began in 2006.\(^{345}\) This one and the initial ClueBot operated with simple text pattern matching.\(^{176}\) Christopher Breneman’s and Cobi Carter’s ClueBot NG, now operating on the English Wikipedia, leaves a major impact by rejecting 40,000 vandalisms a month (mid-2012 number).\(^{345}\) It monitors the IRC stream of changes and uses a trained machine learning algorithm (an artificial neural network) to classify and will then revert the edit if it is sufficiently probable that the edit is a vandalism. Training set data for ClueBot NG’s machine learning is obtained through a web-based interface at review.cluebot.cluenet.org where Wikipedians can label data. Besides the ClueBot NG, other anti-vandalism and article quality prediction systems using machine learning or other complex algorithms have been suggested.\(^{178,205,341–343,346–351}\) Some of these systems do not necessarily focus on vandalism detection, but on more general quality aspects, e.g., classifying “featured” articles from other kinds of articles, see Table 8. Reported in 2006 one system used dynamic Bayesian networks and the BUGS software to compute the “trust” of English Wikipedia articles based the trust of its previous revision, the user role and the amount of insertion and deletions with these variables linked up in a Markov chain.\(^{341}\) The researchers used beta distributions setting priors based on the status of the user, regarding administrators as the most trustworthy and then in decreasing order: registered users, anonymous users and blocked users. The researchers in this particular study restricted themselves to 50 featured articles, 50 “clean-up” articles and 768 normal articles in geography for the computation of trust. It is unclear whether the method using BUGS would scale to full Wikipedia application. Besides computing trust the researchers also used the trust values to predict the article class. On a test set of 200 new articles with 48805 revisions they could report predicting 82% featured and 84% clean-up article correctly.

An anti-vandalism approach from 2008 used bag-of-words and a naïve Bayes classifier as well as ‘probabilistic sequency modeling’, though could not report better than ClueBot results.\(^{178}\) Another approach used features such as ‘edits per user’ and ‘size ratio’ and obtained quite good results.\(^{177}\) The UC Santa Cruz WikiTrust, presented at the Web at wikitrust.soe.ucsc.edu, has an associated API available, which may report the author of a word, its “trust” and the revision where the word was inserted. This API is, e.g., used by STiki. WikiTrust extracts a number of features and uses a classifier from the Waik toolkit.\(^{350}\) The system enables individual words in Wikipedia articles to be colored according to computed ‘trust’.\(^{349}\)

In contrast to the complex algorithm approaches, it has been found that the word count of an article performs surprisingly well as a predictor for article quality.\(^{342,352}\) Blumenstock suggested to make a cut at 2000 words between featured and random articles and found a error rate on around 96% on a corpus of 1554 featured and 9513 randomly selected articles.\(^{342}\) The researcher of this study also tried a number of other more complex features, e.g., readability indices, citation count, category count and sentence count, and a range of machine learning algorithms, increasing the accuracy to 98%. When featured articles are compared to other selections of articles rather than random ones then quality-associated features may change. Using an analysis of the 2007 Virginia Tech Massacre article for hy-
By considering good-quality “tokens” as those edits inserted by a user that are present after the intervention of an admin Javanmardi, Lopes and Baldi reported in 2010 about the performance of models for user reputations.\textsuperscript{205} They examined four different user roles: Admins, “vandals”, “good users” and blocked users. Looking at the entire revision history of the English Wikipedia and utilizing a diff algorithm and MD5 checksums they were able to classify between admins and vandals with an area under the ROC curve value of around 97.5\%. In one of their models the speed of the deletion affected the reputation of the user negatively and in their most elaborate model the reputation of the user that deleted the content also affected the reputation of the user inserting the content.

In 2010 a Wikipedia vandalism classification competition was held, where 9 systems competed. The competition was based on a data set with 32'452 edits on 28'468 different English Wikipedia articles annotated with crowd-sourcing through Amazon’s Mechanical Turk.\textsuperscript{355} A plentora of different features were suggested,\textsuperscript{354} see also Figure 9 for a few. The system by Mola Velasco won with WikiTrust coming in 2nd or 3rd (depending on the measure used for evaluation).\textsuperscript{354} Mola Velasco used a broad range of features and for the classification he used algorithms from the Weka framework.\textsuperscript{351} However, the organizers of the competition could report (with some measures) better performance than any of the individual classifiers when the classifiers were combined into an ensemble classifier.\textsuperscript{354} The competition the following year would also include German and Spanish Wikipedia edits.\textsuperscript{356} The winning entry used 65 features. Among the presently best performing algorithms are apparently a system using 66 features and Lasso for estimation.\textsuperscript{357}

To avoid that good edits automatically get reverted (false positives) the bots usually operate with a high threshold leaving false negatives. This problem has left room for the semi-automated vandal fighter tool of Andrew West called STiki. The original system hooked on to the IRC and the Wikipedia API. With feature extraction and machine learning applying support vector regression running on a server, edits would be scored for vandalism and high-scoring edits would be presented to a client where a Wikipedia could indicate whether the edit was a vandalism or not.\textsuperscript{179, 180} The Wikipedians using the system would need to download the Java client program. The newest version of STiki uses information from ClueBot NG and WikiTrust.

An different approach, Wikibu, seeks to indicate the reliability of a Wikipedia article by embedding the Wikipedia articles on a page that also shows the number of visitors, number of editors, number of links and sources of the Wikipedia article. A version for the German Wikipedia is available from http://www.wikibu.ch.

Other efforts focus on approval of article through peer review. There are already mechanism for peer review on Wikipedia itself (e.g., the notion of ‘featured article’), but also combinations with external systems have been suggeted.\textsuperscript{358} One example was Veropedia, that had over 5000 checked Wikipedia-based articles. In July 2010 the site was unavailable.

Back in 2006 Bertrand Meyer called for a certification of an entry and a refereeing process and believed that a rating system for authors and entries is inevitable.\textsuperscript{167}

**Automatic creation of content**

Andrew Lih has called the start of the mass-creation of content via bots “the most controversial move in the history of Wikipedia”.\textsuperscript{25} In the autumn of 2002 Derek Ramsey began to add articles on US counties to the English Wikipedia based on information extracted from public United States Census data. The automatically constructed text, corresponding to a couple of thousands of articles, he manually added to Wikipedia. After this addition he started with automated creation of articles for 33'832 US cities completing the task in less than a week in October 2002. As the English Wikipedia had just over 50'000 articles before Ramsey began, he quickly became responsible for the creation of 40\% of the articles on the English Wikipedia.\textsuperscript{25} Longitudinal studies show the effect of Ramsey’s work.\textsuperscript{283, 284} Before Rambot’s activity a few other bots operated on Wikipedia creating content on a more limited scale: The ‘Wikipedia:History of Wikipedia bots’ English Wikipedia page lists, e.g., import of a glossary of telecommunication terms in February 2002 from Federal Standard 1037C. Bots can still make a major impact on the size of a Wikipedia: Sverker Jo-
<table>
<thead>
<tr>
<th>Edit feature</th>
<th>Description/assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit time of day</td>
<td>180</td>
</tr>
<tr>
<td>Edit day-of-week</td>
<td>180</td>
</tr>
<tr>
<td>Registration time</td>
<td>Time since registration of editor^{180}</td>
</tr>
<tr>
<td>Last edit time</td>
<td>Time since last edit of article^{180}</td>
</tr>
<tr>
<td>Vandalized time</td>
<td>Time since editor last vandalized^{180}</td>
</tr>
<tr>
<td>Article reputation</td>
<td>180</td>
</tr>
<tr>
<td>Editor reputation</td>
<td>180</td>
</tr>
<tr>
<td>Categorical reputation</td>
<td>180</td>
</tr>
<tr>
<td>Geographical reputation</td>
<td>180</td>
</tr>
<tr>
<td>Comment length</td>
<td>Length of the revision comment^{180}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article features</th>
<th>Description/assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of edits/rigor</td>
<td>More edits are better.\textsuperscript{283,284} “[...] more editing cycles on an article provides for a deeper treatment of the subject or more scrutiny of the content.”\textsuperscript{111}</td>
</tr>
<tr>
<td>Number of editors/diversity</td>
<td>More edits are better.\textsuperscript{283,284} “With more editors, there are more voices and different points of view for a given subject”\textsuperscript{111}</td>
</tr>
<tr>
<td>Word count</td>
<td>The length of an article. Long articles are better.\textsuperscript{342,352}</td>
</tr>
<tr>
<td>Semantic stability</td>
<td>Articles with semantically stable revisions are better.\textsuperscript{353}</td>
</tr>
</tbody>
</table>

Table 9: Quality features of an edit or article. See also the many more features used in the Wikipedia vandalism detection competition.\textsuperscript{354}

hansson’s bot operating on Swedish, Waray-Waray and Cebuano Wikipedias has constructed 2.7 million articles\textsuperscript{xxxi} as of February 2014. The large number of articles for species has made the Swedish Wikipedia one of the largest Wikipedias.\textsuperscript{131}

Scientists have also automatically added information to Wikipedia by formatting data in already existing scientific databases for inclusion in Wikipedia. In the molecular biology project, Gene Wiki, the Protein Box Bot has since August 2007 added several thousand articles on genes,\textsuperscript{240,359} with automated construction of an infobox, free-text summary and relevant publication aggregated from Entrez Gene and a gene atlas database.\textsuperscript{360} In a similar project, the WikiProject RNA started out with the Rfam database containing information about RNA and added the information to Wikipedia, creating over 600 new articles. After the addition of content to Wikipedia the information flow was turned around and the (possibly extended) content of Wikipedia was used for the annotation of the Rfam database by daily monitoring Wikipedia for changes.\textsuperscript{361} Since November 2008 the Rfambot has been working with updates and version of infoboxes in Wikipedia articles related to the project.

Tools may automatically create on-the-fly text from the structured information in Wikidata. Magnus Manske’s Reasonator Web service\textsuperscript{xxxviii} shows one of the first examples, see Figure 12 for the generated page of Johann Sebastian Bach. A related on-the-fly construction of Wikidata semantic information is provided by Metaphacts with https://wikidata.metaphacts.com.

Stuart Geiger has made a longitudinal overview of overall bot and assisted editing activity on the English Wikipedia for an around two month period in the beginning of 2009. Assisted editing amounted to approximately 12% of all edits.\textsuperscript{181,210} Most of these edits seem not to be related to creation of content but rather vandal fighting, but bots can have a major impact on some areas of Wikipedia. With longitudinal examination of the number of scientific citations in Wikipedia I found a big jump from 2007 to 2008 due to content with citations added by the Protein Box Bot.\textsuperscript{81} The WikiOpener MediaWiki extension creates a framework for querying external databases such that their material can be merged in the wiki.\textsuperscript{362} In May 2014 Magnus Manske introduced gamifi-
Figure 12: Magnus Manske’s Reasonator Web service with on-the-fly generation from data in Wikidata. The text in the grey field is template text filled with data from Wikidata.

citation of Wikidata entry, letting Wikidata editors set values of properties (e.g., the gender of a human or whether an item ‘is a’ human) based on the introductory paragraphs of Wikipedia pages. The interface would present buttons for the user to press and the Web service would perform the actual Wikidata editing behind the scene. Google developers constructed another semi-automated Wikidata entry tool, Primary Sources Tool, initially used to move data from Google’s Freebase to Wikidata. Via a gadget the tool would suggest new claims by modifying the Wikidata edit interface and giving the user a choice to either accept or reject the suggested claim. In January 2016 users had performed about 90,000 approval or rejection actions.

Automatic creation of links

In a procedure what may be called automated link discovery tools suggest intrawiki links from a word in a Wikipedia article to an appropriate Wikipedia article. At one point Wikipedia user Nickj ran the Link Suggester and Can We Link It. Another user ran a link discover bot from the account Anchor Link Bot. The DPL bot does not suggest new links but notify (human) editors about links they have added, that points to disambiguation articles and may need to be resolved.

Work suggestion

A standard installation of the MediaWiki software can generate lists for work suggestions. MediaWiki users find these links on the page ‘Special:SpecialPages’ and examples are ‘dead-end pages’, ‘wanted categories’ and ‘wanted pages’.

The Wikidata-based service Wiki ShootMe! lists geographical items with missing images on Wikidata based on a query coordinate given by the user. A Wikipedian can use it to identify photo opportunities in his/her nearby area. The MediaWiki software embeds similar functionality with the ‘Special:Nearby’ page that lists nearby pages and associated images if the use shares his/her estimated browser geolocation.

Semantic wikis

The framework referred to as the ‘Semantic Web’ structures data. Essentially, a subject-verb-object triple data structure represents data, where the triple data may be represented in RDF. In the middle of the 00s several research groups proposed systems that merged the Semantic Web idea with wiki technology with some of the systems based on extensions around the MediaWiki software. Researchers compared 10 different semantic wikis in a 2008 paper and ended up selecting the Semantic MediaWiki system for their knowledgebase.
In the German Semantic MediaWiki (SMW) the wiki page determines the subject of the triple structure, while the wikilink sets the object. The verb of the triple structure comes as an element in an double colon extension to the MediaWiki link syntax, so that, e.g., \([\text{has capital::Berlin}]\) on the Germany page will result in the triple structure (Germany, has capital, Berlin). In the Japanese Semantic MediaWiki\textsuperscript{377} the suggested syntax would be \([\text{Term:Berlin has capital}]\). The German SMW comes with query facility, so user may formulate semantic on-the-fly queries across all the semantic data of the wiki.

Though not implemented in Wikipedia nor its sister projects, the SMW extension has proven valuable in many other contexts, and a large number of extension on top of SMW have come into existence,\textsuperscript{28} e.g., the Semantic Result Forms extension allows user to let the semantic query return the result in a large number of formats, both text-oriented and in plots of different kinds. The SemanticHistory extension can enable semantic queries on the wiki edit history.\textsuperscript{381}

The Semantic Web does not directly allow for specification of n-ary relations beyond the triplet structure such as (Germany, has capital, Bonn, in year, 1990). However, an n-ary relation can be broken down into multiple triplets.\textsuperscript{382} The BOWiki extends Semantic MediaWiki so n-ary relations can be represented with the wiki syntax.\textsuperscript{383} Extensions for Semantic MediaWiki also allows for representation of n-ary objects.

## Wikidata

Within the MediaWiki framework the 2012 Wikidata proposal\textsuperscript{385} discussed the issues and presented a mockup with field/value editing.\textsuperscript{384} This system became operational in October 2012, and major bot activity soon made it the largest Wikimedia wiki in terms of pages (or ‘items’ in the parlance of Wikidata) and in terms of editing activity with about 90% of the edits made by bots.\textsuperscript{385} Inspired from the Semantic Web and semantic wikis the page of Wikidata acts as the subject, while the verb is referred to a ‘property’ and the object will refer to another Wikidata page or another type of value. Wikidata allows n-ary relations through so-called qualifiers. Limited support for units, such as heights and mass, came in September 2015.

The query language on the base installation of Wikidata was initially limited with no means to build advanced queries like the ones seen in SPARQL or Freebase’s MQL. However, a few third-party services soon appeared with Magnus Manske as the primary driver behind the tools. His Wikidata Query editor at \url{http://wdq.wmflabs.org/wdq/} allowed for formulation of queries such as “Places in the U.S. that are named after Francis of Assisi”, and his autolist tool at \url{https://tools.wmflabs.org/autolist/} would display a list of retrieved pages. Soon Kingsley Idehen’s OpenLink Software company also provided a Virtuoso-based SPARQL endpoint for querying Wikidata together with other Semantic Web data from the Linked Open Data cloud.\textsuperscript{xlii} Metaphacts and University of Chile also made early public SPARQL endpoints with Wikidata data available.\textsuperscript{386} Finally, in September 2015 Wikimedia Foundation announced its SPARQL end-point, the Wikidata Query Service, running from \url{http://query.wikidata.org}.

With its qualifiers and references the Wikidata data structure does not map directly to the triple structure of the Semantic Web and SPARQL tools. However, various approaches—reifications—can convert Wikidata data to the triple format. Benchmark profiling with four different approaches and five different SPARQL engines found that the so-called n-ary relation model enabled the expression of property paths.\textsuperscript{387}

When Wikidata is coupled with natural language question parsing powerful open-domain question-answering (QA) systems can be constructed. Students from École Normale Supérieure de Lyon constructed one of the first online QA systems with Wikidata data, the Platypus system running from \url{http://askplatyp.us}.\textsuperscript{388} It was capable of answering questions such as “Who is the prime minister of France?” and “Who was member of the Beatles?”.

### Other structured content

Several systems exist that extending wikis with table and database-like functionality beyond the Semantic Web framework, e.g., DynaTable,\textsuperscript{xiii} that defines structured data in a separate name space and can display the structured data in tables on wiki pages with the use of special tags.\textsuperscript{389} Other similar extensions with varying degree of maturity are WikiDB and DataTable extensions. The relatively simple TemplateTable extension extracts keys and values from templates used across the wiki and generates a table with columns according to key and each row corresponding to a template instantiation. No filtering is possible. The

\textsuperscript{xii} \url{http://sourceforge.net/projects/wikidynatable/}

\textsuperscript{381} \url{http://www.mediawiki.org/wiki/Wikidata}

\textsuperscript{xiii} \url{http://lod.openlinkkw.com/sparql}

\textsuperscript{xlii} \url{http://askplatyp.us}
Brede Wiki keeps its MediaWiki template data in a sufficiently simple format for complete databasing the template content by extraction from XML dumps, enabling complex SQL data mining queries to be made off-wiki. The Brede Wiki also keeps table-like data in simple comma-separated values format on wikipages for meta-analysis. The SimpleTable extension formats the CSV data.

None of these systems enjoy as large a success as the semantic wikis and Wikidata.

Form-based editing

Standard wikis have only one input field for the content of the Wiki. Semantic Web-oriented wiki engines, such as OntoWiki, present edit interfaces with one input field for each property of a page, thus each pages will typically have many forms. The Swiki system can associate forms to wiki pages and the template mechanism in the CURE wiki engine allows the editor to define the interface for editing a multiple input form as well as the presentation of the content.

An extension for the MediaWiki software brings support for editor-defined forms, used in conjunction with Semantic MediaWiki and the template functionality. As with Semantic MediaWiki the form extension is not enabled in Wikipedia, but is used on a number of SMW-enabled wikis ensuring a consistent input of data.

With Wikidata Wikipedians got the opportunity to make form-based edits on language links and structured content as usually found in infoboxes. Wikidata form-based input helps the user by suggestions for autocompletion.

Markup

Wikis have a simple markup language. Unfortunately the markup have diverged slightly so the markup is not entirely compatible between wiki engines, e.g., MediaWiki does not support CamelCase linking. WikiCreole (or just Creole) attempts to standardize markup and some wikis have adopted it. see http://www.wikicreole.org/wiki/Creole1.0. The WikiCreole language has also been formalized in ENBF.

The difficulties with representing MediaWiki wikitext so standard context-free grammars may parse it means that creating tools for WYSIWYG editing is difficult. Software developers within the Wikimedia Foundation project Parsoid implement a node.js program that may represent MediaWiki wikitext as HTML5, and convert the HTML5 back to wikitext. The project should not only make enabled good visual editing in MediaWiki-based wikis, but also increase rendering performance by avoiding conversion overhead.

Text-oriented formatting languages, such as E\LaTeX and HTML, can have fairly dense markup obscuring the text itself. Documentation languages focus on readability while maintaining formatting capabilities, e.g., AsciiDoc, Markdown and reStructuredText. These language typically differ from wiki markup, though, e.g., GitHub Flavored Markdown has URL autolinking like typical wiki markup.

Extended authoring

The interface and the wiki markup may not appeal to potential contributors with low level of technical ability. If they are knowledgeable contributors, Wikipedia have lost the benefit of their knowledge. In 2008 the Wikimedia Foundation received a grant from the U.S.-based Stanton Foundation for a project to make the editing facility easier.

Years later in 2013 the Wikimedia Foundation rolled out the “WYSIWYG-like” editor VisualEditor on Wikipedia giving contributors the choice between usually raw wiki markup and WYSIWYG-like editing. Editing defaulting to VisualEditor generated controversy among Wikipedians, and English Wikipedia administrators manage to revert the decision making the VisualEditor an opt-in option.

Some wikifarm companies, such as wetpaint and PBwiki, have WYSIWYG functionality. MediaWiki-based Wikia wikifarm has WYSIWYG support. On Wikipedia user Cacycle’s wikEd Javascript in-browser text editor adds enhanced text processing.

The Firefox web-browser plugin Zotero formats bibliographic references in the style of Wikipedia citation templates. ProveIt from Georgia Tech (http://proveit.cc.gatech.edu/) provides a structured interface for editing references in MediaWikis via a popup window.

Digitizing publications

Wikimedia Foundation project Wikisource collects electronic versions of free content publications. Project Gutenberg (http://www.gutenberg.org/) started by Michael Hart has a longer history going back to 1971 and had by 1992 digitized many works. Lars Aronsson’s Project Runeberg is a Scandinavian effort inspired by Project Gutenberg.
To solve conflicts about differences between editions Aronsson started in 1998 to incorporate facsimiles with optical character recognition (OCR) and collaborative proof-reading. Aronsson’s Runeberg approach has now been adopted by Wikisource. MediaWiki still forms the basis for Wikisource, but enables the Proofread Page MediaWiki extension and has a ‘Page’ namespace where a facsimile page is presented next to the wiki edit field containing OCR’d text from PDF or djvu files for collaborative proof-reading. The approach splits the content and presentation on different wikipages and a full book may be split across a number of pages, so tools using Wikisource as a corpus may need to implement a not straightforward procedure for assembling of the actual texts and discarding of wiki markup.

Facsimiles of paper encyclopedias may have multiple articles on one page or one article spanning multiple pages, so facsimiles present two structures: the page-oriented structure and the chapter/article-oriented structure. In Wikisource and Runeberg proof-reading is page-oriented while the presentation of proof-read books is chapter-oriented in Wikisource.

Geographical extension

Wikipedia lets users markup a page with geographical coordinates via the use of MediaWiki templates, — a simple approach to so-called volunteered geographic information (VGI). Other wiki-oriented systems present a map that users draw on while also providing descriptions of the geographical coordinates. One such system is WikiMapia (http://wikimapia.org) where users can annotate with rectangles and polygons on a background of satellite images from Google Maps. The CC-BY-NC-SA licensed data are exposed through an API.

One of the most (if not the most) prominent VGI efforts is OpenStreetMap (OSM), www.openstreetmap.org, that also provides complete maps under share-alike license. Founded by Steve Coast in July 2004, users may add, e.g., roads and nodes with key-value attributes with the help of a background satellite image and GPS tracks. Rendering may be through the Mapnik open source library. OSM content may be integrated with Wikipedia, so articles with geographical coordinates display OSM maps, e.g., through the OpenStreetMap SimpleMap MediaWiki extension.

Another geographically oriented wiki is WikiCrimes (http://www.wikicrimes.org) that seeks to map crimes collaboratively.

Extending browsing

A number of tools exists for reformating Wikipedia content for specialized devices. Jesse David Hollington has reviewed applications for the iPhone, including the Wikipanion with features such as cache and offline reading.

Wikipedia Diver is a Firefox plugin. It logs clicks between Wikipedia pages to a database and displays the browsing path graphically. The Wikipedia Beautifier Chrome extension presents a cleaner view of the text with automatic hyphenation and various adjustments to the Wikipedia interface. Googlepedia Firefox add-on shows relevant Wikipedia articles next to Google search results. A number of other browser plugins changes or enhance the presentation of Wikipedia in the browser.

Graphic extensions

The image-oriented browsing program Indywiki enables a user to search Wikipedia independently of an ordinary Web browser. The program displays the ten most related images prominently in its browser interface while also displaying the text and links of the article. A similar system is the Qwiki website that display images from Wikipedia and other sources in a multimedia environment.

The Java program WikiStory constructs interactive time lines based on Wikipedia material. The Web application HistoryViz displays events related to a queried person on a timeline. Apart from this visualization the system also features a Java applet graph visualization of Wikipedia pages. The system relied on algorithms for categorization of Wikipedia articles into persons, places or organizations. Yet another timeline visualizer is Navino Evans’ histropedia served from histropedia.com. Timeline visualization features also as part of the Reasonator Wikidata presentation Web service.

Wikis can be extended with on-the-fly creation of graphs and there exists an extension for MediaWiki that use the GraphViz software, and, e.g., the Brede Wiki uses the tool for automatic creation of visualizations of intrawiki links between brain regions, topics and organizations from information defined in MediaWiki templates. The intrawiki links between Wikipedia articles form a directed graph. Together with a path finding algo-

---

xlviihttp://indywiki.sourceforge.net/
Algorithms Erika Arnold’s wikiGraph Web service can find and visualize a path between two user-specified Wikipedia articles using the dump and the Neo4j graph database.\textsuperscript{4,58}

The ‘Copernicus’ system makes an attempt on a 3D wiki:\textsuperscript{410–412} A two-layer interface presents the Wikipedia article in a transparent foreground, while the background presents a 3D model related to the Wikipedia article. The user can trigger predefined camera movements and adjust the transparency.

Video extensions

Multimedia authors have added animated images and small video clips in the Ogg Theora format in Wikipedia. Other editors can in principle edit these clips with external video editing software, but with far less ease of use as when editing wikitext. The company Kaltura have worked on a Open Source technology to better support peer production of video content. In the beginning of 2008 the company announced experiments with this collaborative video platform together with the Wikimedia Foundation. As with other material Wikipedia already enables collaborative tagging and discussion of videos through category markup and discussion pages on Wikimedia Commons.

Real-time editing

When two editors edit the same time on a MediaWiki article and each one saves his/her edits then the editor with the last save will experience an edit conflict that needs to be resolved manually. In this aspect MediaWiki differs not from standard revision control systems Concurrent Versions System (CVS) and Subversion (SVN). Newer software revision control systems, such as Git, may be distributed. There have been several suggestions and developments for distributed wikis that merged the wiki idea with the distributed revision control system idea. Levitation would convert a Wikipedia database dump to Git repositories. In Ward Cunningham’s Smallest Federated Wiki a user will fork a wiki when he edits. The owner of the original edit may then merge that edit back to the original wiki.\textsuperscript{417} Some of his thoughts on federating wikis go all the way back to a 1997 memorandum.\textsuperscript{418} Another distributed wiki system is research prototype Concerto (http://concerto.xwiki.com), that was built for the XWiki wiki engine as a peer-to-peer replicated wiki with an aim for data consistency between the replicated wikis.\textsuperscript{419}

Distributed wikis may store content that diverges from neutral point of view: What has been called “every point of view”.\textsuperscript{420} The question on whether distributed wikis (Wikipedias) might also help for an inclusionist approach to specialized knowledge, have also been posed.\textsuperscript{421}

Distributed and disconnected Wikipedia

Wikipedia can be downloaded and read from the local copy. Groups of editors have checked a selected number of articles and distributed them on DVDs, see www.wikipediaondvd.com and Schools-wikipedia.org. The offline multimedia reader Kiwix makes Wikipedia available offline. The simple small AAA battery-powered WikiReader comes with touchscreen and 3 million topics from Wikipedia. An even simpler device, Humane Reader, connects to TV and could include an offline version of Wikipedia.\textsuperscript{414} The One Laptop Per Child (OLPC) project features the Wikibrowser with a selection from the Spanish and English Wikipedia, while the XOWA Java tool enables the setup of local offline copies of Wikipedias and their sister projects as well as their images.\textsuperscript{415}

Offline Wikipedia deployment may be important in areas with no Internet, and researchers have studied the use of Wikipedia in offline OLPC in this context. In a field study Ethiopia students indicated that “browsing schools books and offline Wikipedia page” was the second most favored activity on the OLPC.\textsuperscript{416}

Versioning in Wikipedia is centralized as in the client-server software revision control systems Concurrent Versions System (CVS) and Subversion (SVN). Newer software revision control systems, such as Git, may be distributed. There have been several suggestions and developments for distributed wikis that merged the wiki idea with the distributed revision control system idea. Levitation would convert a Wikipedia database dump to Git repositories. In Ward Cunningham’s Smallest Federated Wiki a user will fork a wiki when he edits. The owner of the original edit may then merge that edit back to the original wiki.\textsuperscript{417} Some of his thoughts on federating wikis go all the way back to a 1997 memorandum.\textsuperscript{418} Another distributed wiki system is research prototype Concerto (http://concerto.xwiki.com), that was built for the XWiki wiki engine as a peer-to-peer replicated wiki with an aim for data consistency between the replicated wikis.\textsuperscript{419}

Distributed wikis may store content that diverges from neutral point of view: What has been called “every point of view”.\textsuperscript{420} The question on whether distributed wikis (Wikipedias) might also help for an inclusionist approach to specialized knowledge, have also been posed.\textsuperscript{421}

Wiki and programming

There are different ways to combine programming and wikis. The extension framework of MediaWiki allows developer to make additional programs that use resources either on the wiki server or on remote servers. Other approaches enables end-user programming for editors, e.g., one system allows editors to write database queries for a SAP system.\textsuperscript{422} A MediaWiki extension integrates output from the \textit{R} statistical computing environment.

Wikis that allow for inclusion of different applications, e.g., calendars, may be termed ‘application
wikis’.

The MediaWiki software has a simple ‘programming language’ for templates, which features, e.g., an ‘if’ sentence. However, the capabilities are limited and the syntax quite obscure with excessive number of curly braces. In 2012 Wikimedia began serious experiments with enabling the Lua programming language on test MediaWiki installations. A successful application would fundamentally change the way that MediaWiki users write templates and make way for more advanced programming. In the first part of 2013 Lua was enabled on the Wikipedias.

A programming language embedded in a wiki could provide means to format data, e.g., comma-separated values into tables, and could be used to make small scale numerical computations, — functions that now needs to be programmed either in a MediaWiki extension or in an external web script.391

The PyPedia website www.pypedia.com combines MediaWiki and the Python programming language. With each wikipage containing a Python function or a Python class PyPedia allows Internet users to write and execute code in the browser or execute the code on the wiki in a local installation of Python.423

5 Using Wikipedia and other wikis in research and education

As a Web-based collaborative environment Wikipedia and other wikis offer researchers and students on all levels a means for communicating, — reading and writing about the topic at interest, e.g., wikis can offer Ph.D. Students an environment for literature research where a wiki page contains information about a specific research paper: its summary, research questions, methods, results and comments.393

In Ten simple rules for editing Wikipedia Logan et al. guide scientists how to write on Wikipedia.424 For example they suggest scientists to register an account for privacy, security and reputation building as well as to gain access to the “watchlist” feature. They suggest that scientist should “avoid shameless self-promotion” by not writing their biography page on Wikipedia (let others do that).

Attitudes towards Wikipedia

A Wikimedia Foundation survey has found researchers to be generally quit positive towards Wikipedia: Over 90% of 1743 self-selected respondents were ‘very favorable’ or ‘somewhat favorable’.425 Among Public Library of Science (PLoS) authors the result was 96%. Other results showed that 68% answered ‘yes, on a large scale’ to the question ‘would you be in favor of effects to invite scientist to add or improve Wikipedia articles’. Such results are very positive for Wikipedia, but may be biased due to the self-selection of respondents and that the publisher web site with initial reference to the survey was Open Access.

In 2010 Dooley surveyed 105 university faculty members and found that 54.4% would rank Wikipedia to be moderately to very credible, 26.6% would said it had some credibility, although not much and 20% that it had “no credibility”.426

Individual researchers have called out for colleagues to contribute to Wikipedia, e.g., “Scientists who receive public or charitable funding should [...] seize the opportunity to make sure that Wikipedia articles are understandable, scientifically accurate, well sourced and up-to-date.”427 In 2011 Alexander Bond called out for ornithologists to appropriate Wikipedia “as a teaching and outreach tool”, e.g., “[P]rofessors can replace essays and reports assigned to students with the creation or improvement of a taxonomic Wikipedia entry” and recommend the WikiProject Birds page as a starting for ornithologists.428 Writing for students and lawyers that would read Wikipedia, Diane Murley concluded that “Wikipedia can be a great research tool if used appropriately” and that “it is a good tool to use to teach researchers about the necessity of evaluating sources”.429

Some scientist contributing to Wikipedia have voiced concern about the priorities of Wikipedia. Antony Williams asked in a blog post in 2011 “Why are pornstars more notable than scientists on Wikipedia?” noting that scientist biographies he had written had been flagged for notability, even though one of the biographed scientists had over 300 publication and an h-index of 27. Williams’ own Conflict of Interest in the case had been questioned. Direct Wikipedia contributions from scientists, as well as other experts, are not automatically regarded as appropriate. Wikipedia is not interested in what experts know, but rather from where the experts know their knowledge, and that the source has “institutional approval”.

Rosenzweig argued that professional historians can even learn from the open and democratic production and distribution model of Wikipedia, and he asked why scholarly journals are behind paywalls and sponsored American National Biography Online is available only to libraries at a high cost.40

He thought that historians should join in writing
history in Wikipedia, but also noted issues with the ban on no original research and “difficult people” that drive experts away as hindrances for professional historians writing on Wikipedia. He called for systems for digitizing historical documents and for open-source textbooks, — puzzlingly not mentioning the Wikisource (with oldest mainpage edits in November 2003) and Wikibooks (with oldest mainpage edits in May 2005) efforts.

Echoing Rosenzweig in another domain, Wikipedian and healthcare professional James Heilman encouraged healthcare professionals to get involved in Wikipedia stating among his reasons that “Wikipedia has taught me critical reading, which has made me better equipped to deal with less reliable sources of information such as pharmaceutical representatives.” The positive attitude towards Wikipedia in the healthcare domain has led to a closer collaboration between Wikipedians and members of the Cochrane Collaboration for strengthening the dissemination of synthesized research.

Use of Wikipedia

Several studies have used surveys to examine the use of Wikipedia among university faculty and students. These studies may aim to answer how much and for what purpose they use Wikipedia. Apart from the direct academic and cognitive purposes of obtaining knowledge and satisfying curiosity other motivations for using Wikipedia may be hypothesized: emotional, personal integrity, social and tension-release needs.

The Wedemeyer study asked students to evaluate Wikipedia biochemistry articles. One third responded that they never use Wikipedia. Among the remaining two thirds 12% used Wikipedia as their primary source and 31% used their textbook and Wikipedia equally. The remaining 57% used Wikipedia only as a supplement. The majority of the students preferred Wikipedia to the textbook.

Sook Lim performed a survey among undergraduate students, where 134 out of 409 responded. Reported in 2009 and 2010 she found that all students used Wikipedia, but that the use was not necessarily for academic purposes alone as students would also use Wikipedia for “entertainment and idle reading” and nonacademic personal information interests. That was especially the case for the male students that also had a higher perception of information quality and belief in the Wikipedia project. Another study reported in 2011 found that 47% of 186 medical students who recently had completed psychiatric clinical clerkship used Wikipedia as one of the primary sources for preparing for psychiatry exams. Question books (88%) and the peer-reviewed website Up-to-Date (59%) were more frequently used, but textbooks (10%) less used. Among the students using Wikipedia 84% also used question books.

In Dooley’s 2010 study among university faculty 45 of 105 respondents said they used Wikipedia in their teaching and/or research, 40 occasionally, and 20 respondents said they never used Wikipedia for teaching/research.

In a paper discussing how students could use Wikipedia Jeff Meahre suggests that students pursue the discussion pages to “see the process of knowledge creation” and show why citations are important.

In a study on high school students’ writing assignment for on a wiki the researchers found that they would use Wikipedia even though they were aware of its limitation as a source.

Citing Wikipedia

Many users would say that Wikipedia works well for background reading and doubt that you can use Wikipedia as a source. In the beginning of 2007 a department at the Middlebury College would hold students responsible for using Wikipedia as a source after a batch of students had used erroneous information on Wikipedia about topics in the history of Japan (Shimabara Rebellion and Ogyu Sorai). Media reports implied that the department of Neil Waters, the teacher of the class, ‘was at war with Wikipedia itself’. However, Waters himself actually told students ‘that Wikipedia is a fine place to search for a paper topic or begin the research process’. The policy adopted by the department was:

1. “Students are responsible for the accuracy of information they provide, and they cannot point to Wikipedia or any similar source that may appear in the future to escape the consequences of errors,

2. Wikipedia is not an acceptable citation, even though it may lead one to a citable source.”

At Lycoming College a teacher in history would outright ban Wikipedia from student’s bibliographies giving a grade of zero if any such citation appeared.

Waters’ policy is in line with the opinion of the Wikimedia Foundation. However, in the end of 2007 Jimmy Wales said that he saw no problem in younger students using Wikipedia as a reference, and that it should be used as a stepping stone to other sources.
It is fairly easy for a teacher to point to correct information on Wikipedia by the use of permanent links to a specific version of an article. Teacher may link to the version s/he has reviewed, so that students may feel comfortable in using the information on Wikipedia and use it as a supplement to textbooks.

Citing Wikipedia will be a problem if administrators delete the page, e.g., due to notability. Both the present and the older versions of the article will be inaccessible to the editor and reader.

Regardless of the problems with citing Wikipedia research papers do indeed cite Wikipedia. Dooley examined 250 research reports published in 2009 and in the beginning of 2010 that returned on a query on Wikipedia from a search on Academic OneFile electronic database: 249 of the papers used Wikipedia as a source of scholarly information. In 27 of the papers Wikipedia was the main topic and 62 had brief mentions of Wikipedia. In a similar citation study Brazzeal would find 370 chemistry article from 2005 to 2009 citing Wikipedia. The author had searched among 312 chemistry journals from the websites of the three publishers Elsevier, Springer and the American Chemical Society. Of these 312 journals 147 had at least one citation to Wikipedia. 9% percent of the articles had a citation to numerical values of physical and chemical properties in Wikipedia. In a 2008 study in the area of humanities Lisa Spiro retrieved 167 results between 2002 and 2008 citing Wikipedia. In 2008 RNA Biology became probably the first journal to venture into this approach, requiring authors, that submitted a paper to a section in the journal, to also submit a Wikipedia page summarizing the work. The first article became A Survey of Nematode SmY RNAs. The journal article would undergo full peer-review and thus act as a source for the Wikipedia content. Another case came in 2012 where the scientific journal PLoS Computational Biology allowed researcher to submit educational “topic articles” that were meant to be both published in the journal as well as included in Wikipedia. The first such article was “Circular Permutations in Proteins”.

It is interesting to see whether Wikipedia articles can compete (and even outcompete) the traditional scientific review article.

Special science wikis

Wikipedia has inspired numerous other science wikis. Bioinformatics alone has already quite a large number of wikis. Several biomedical researcher have suggested dedicated wikis for their field since around 2006. The reason why scientists choose to operate their own wiki instead of adding information to Wikipedia is probably due to one or more of four issues:

- The control and ownership of data (an issue particularly important for companies),
- the addition of specialized interfaces and data,
• the problem of notability in Wikipedia, i.e., certain parts of the research information is viewed to be too specialized for a general encyclopedia,

• the control of authorship

Several wikis are up and running handling research information: A Leipzig group extended the MediaWiki software incorporating semantic wiki functionality for gene function (http://bowiki.net). Other groups have created wikis for laboratory protocols (OpenWare), single nucleotide polymorphisms (the SNPedia), and biological pathways (WikiPathways). Other Internet-based biomedical databases have added community annotation, e.g., CutDB, CHDwiki for congenital heart defects and YTPdb for classification of the yeast membrane transporter use the WikiOpener extension to include material from bioinformatics Web databases in the wiki. Other bioinformatics wikis are ArrayWiki, BioData Mining, EcoliWiki, Open Notebook Science, PDBWiki, PiQSi, Proteopedia, TOPSAN, WikiGenes, and WikiProteins. A large share of the scientific wikis uses the MediaWiki software, and often with extensions for handling structured data, e.g., OmegaWiki for WikiProteins or Semantic MediaWiki for SNPedia.

The Rfam database for RNA families is an example of a combination of scientific expert curated database and wiki. The summary information of Rfam was moved to Wikipedia and now the free-form summary information in Rfam is included from Wikipedia with the result that “all of the family annotations are now greatly improved”. The Rfam database edit button leads to the Wikipedia edit page. Rfam-specialized data in Rfam “remain curated by Rfam’s specialist annotators”. With the added information in Wikipedia around 15% of web traffic to the Rfam database was in 2010 driven by Wikipedia.

The problem of notability was the reason I started the Brede Wiki. Wiki pages on this wiki may each describe an individual research article and usually such articles will not be deemed notable enough for Wikipedia as a general encyclopedia. A specific idea of Brede Wiki was also to maintain structured information in a machine readable format for data mining. Related wikis summarizing academic work are Acawiki, WikiPapers and WikiLit, all using Semantic MediaWiki.

Scholarpedia is a MediaWiki-based science wiki especially strong on computational neuroscience and related areas and has strong control over authorship. Operated with Eugene M. Izhikevich as Editor-in-Chief it lets invited experts write specific articles and let other experts review the article. He has had the ability to attract Nobel prize winners as authors. At least one of the articles are archived in the PubMed Central and reference in PubMed under Scholarpedia J.

Other wikis also control the authorship: Gunfgyd and AskDrWiki limit the editing to registered medical doctors and the CC-BY-SA MediaWiki-based Mediapedia similarly controls the authorship by requiring editors to be either approved physicians or scientist with doctoral-degree in biomedical specialities. Finally the general encyclopedia Citizenium does not allow anonymity.

Censorship

In 2008 Bryan Derksen uploaded images of Rorschach test inkblots to the Commons, and in June 2009 James Heilman added the complete set of images to the Rorschach test article on the English Wikipedia. Along with the images came most common response to the each inkblots. Psychologists administring the test angrily held that publishing such information on Wikipedia would jeopardize this psychological test, and American Psychological Association executive director of science Steve J. Breckler said “We wouldn’t be in favor of putting the plates out where anyone can get hold of them”.

Some amount of censoring may occur for certain ‘dangerous’ information, such as the manufacturing of acetone peroxide or details of Zippe-type centrifuge and gas centrifuge.

Personality rights may vary between countries, weighing differently the border between free speech and the right to private life. The Wikimedia blog reported that a German university professor brought suit against the Wikimedia Foundation. He felt that the German Wikipedia violated his right of personality, particularly with respect the mentioning of his membership of student associations. The report did not mention why the professor felt this way. To me the membership seems

Scholarpedia change announced on the Connectionist mailinglist.
entirely innocuous. However, the mentioning of it could be disputed. Denmark treat personal information, e.g., concerning race, ethnicity, religious, philosophical or union affiliations as well as health, sex and gender in a special way. Data cannot usually be registered and distributed without consent, and the mentioning of his membership could expose his religious affiliation. However, in the case with German professor the Wikimedia blog stated that “the information at issue was both accurate and freely available on several other websites under Asche’s authorization”, and the Wikimedia Foundation won the case.466

Scientists on Wikipedia

Elvebakk in her analysis of online philosophy resources speculated that philosophers have added Wikipedia entries for themselves as a form of self-promotion.83

Carl Hewitt, a retired professor from MIT, came into the spotlight at the end of 2007 after edits on the Wikipedia.467 He has a long career within programming, but though he has made significant contribution within the area, several other Wikipedians have regarded his edits as self-promotion. Among others, he have added description of his research (the so-called actor model) in physics articles, where others thought they did not belong. The disagreement led to long discussions and a case for the Arbitration Committee, which decided to ban him.

The case shows an example on how difficult it may be to write from a neutral point of view if there is a conflict of interest and one is deeply engaged in a field of research. However, the guidelines of Wikipedia allow specialists to add their own research, if has background in a reliable publication and follows other rules set by Wikipedia.81

I have myself added a few references to articles I have authored. Universities expect researchers to make their work more widely known, and extending Wikipedia is one way to spread research both fellow researchers as well as ordinary information seekers.

Researchers may be quite attentive to what is written on Wikipedia about themselves: In a 2011 Nature poll, recruiting 840 researchers through email and social web sites, 19% responded that they once per week or more often check Wikipedia for citations to themselves or their own work. The corresponding number for citation-counting sites (such as ISI Web of Science) and search engines were 30% and 38%, respectively. The same poll reported that 9% of the researchers had edited or inserted a reference to their work on Wikipedia within the last 12 months and around 3% has edited their own Wikipedia biography.468, 469

Wikipedia and wikis as a teaching tool

Although educators generally warn against using Wikipedia as a authoritative resource, some promote Wikipedia as an example of a digital environment with rich opportunities for collaborative writing and for considering “new roles and relationships for writers”.470 In a widespread Wikipedia engagement teachers give writing assignment to students in Wikipedia editing,471 e.g., at the University of East Anglia students of International Relations should edit and write article about the Middle East.472 They were rated based on their improvements of eight articles and should write one of their one. The often very controversial topic in this area demanded that the students were able to balance the different opinions and support with sources. Similar assignments were given in history in Lycoming College,437 “Exploring the Past” course at the University of Baltimore473 and environmental history at the University of Washington-Bothell.474 Some teachers choose to use dedicated wikis, such as CoWeb (a technology that goes back to 1998)211, 212 or an installation of MediaWiki.434

Apart from Wikipedia writing assignments, teachers may also give assignments in Wikipedia article review, adding references to an already existing article, translating or copy editing.471, 475, 476 Yet another exercise has students monitoring the changes made to the article they have contributed to and post their findings.437

Students benefit from the Wikipedia writing assignments by having their work exposed not only to the teacher but to the entire world.471 Students also mention gaining skill of doing thorough research as a benefit.475 Stronger research base is also mentioned by teachers.437 In wikis, interaction between peer students may help to improve their writing and in an open wiki the students may become aware of the audience112 and contemplate on the value of authority.437 Wikis may also make the writing process more visible though students revise extensively offline.212 Furthermore, students using wikis may get a better attitude towards collaboration and higher grades.471 Teachers mentions other advantages such as student improvements in computer literacy, Internet resource critique and collaborative work preparation.476 Student reaction to the assignment may initially be worry and anxiety and later reactions may be irritation, pride and indignation.437

---

81 See “Wikipedia:Conflict of interest” (388727443) and “Wikipedia:No original research” (388842869).
In a graduate seminar on plant–animal interactions participants assessed the quality and content of ecology content on Wikipedia. They found Wikipedia generally limited in depth and breadth and with too few citations. Then they proceeded to edit Wikipedia in their domain and found the process “straightforward and efficient, particularly once we learned the protocol for proposing and implementing changes”.

Apart from having the students learn the topic and the teacher having a means to evaluate the students, Wikipedia assignments may also help to construct learning material when no good textbook fully covers the area.

For smaller Wikipedias the work of the students on Wikipedia may not catch sufficient attention from other Wikipedians. In a small study on the Danish Wikipedia historian Bekker-Nielsen monitored 17 Wikipedia articles in 380 days after his students had created the initial articles, e.g., for Demetrios Poliorketes and Carausius. The highest number of edits an article received in the 380 days was 26. Of the total number of edits only 6% were content-related. Bekker-Nielsen also pointed to one worsening edit by a Wikipedian as well as wrong references in one article created by the student and not removed by a Wikipedian before Bekker-Nielsen’s study was made public.

Among the problems faced in Wikipedia assignments are, e.g., student username being censored and students banned because of repeated insertion of copyrighted material. If students work on controversial and highly visible topics the article may suffer vandalism and it may be difficult to “disentangle” the contribution of the student. Combination of little notability and poor writing may result in the student article gets deleted. In a writing assignment on history a teacher found that the vast majority of articles created by students were deleted. Modern educational software systems often use automated plagiarizing detection software. If the student submit directly to Wikipedia then the teacher may have difficulty in evaluating the extend (if any) of plagiarism, and as Wikipedia articles get copied to numerous sites (and possible into the plagiarizing detection software) it may quickly become difficult to track down the origin of a text. Teachers can get around the plagiarizing detection problem by having the students submit their Wikipedia assignment to the teacher before any Wikipedia edits are made.

Descriptions of courses that used Wikipedia as a teaching tool appear on the page Wikipedia:School and university projects. In 2011 Klaus Wannemacher would review these projects. He found 132 projects on the English Wikipedia and a number of others on other language versions of Wikipedia. The earliest dated all the way back to 2002. The field associated with most projects were humanities followed by social science, engineering, medical and natural sciences. He noted teacher issues: making student understand the implication of the license under which they would publish their work, choosing Wikipedia text and the construction of warm-up tasks (creating an account, basic editing, wiki syntax experimentation and categorization). Aims mentioned were increase in student motivation and knowledge, learning to do collaborative writing and gaining the “rigour and balance in writing encyclopaedic articles” with neutral point of view, theoretical analysis of Wikipedia and propaedeutic: making research, editing and bibliographic processing. Some projects attempted to reach "Good Article" Wikipedia status or better while also failures occurred: In one case only 7 of 70 produced articles survived.


Using wikis for course communication

Teachers may use wikis for asynchronous communications in a course. As an encyclopedia Wikipedia should not contain information about specific courses or discussion between individual teachers and students. Another Wikimedia Foundation project, Wikiversity, serves the purpose of building a resource for teaching and learning. A quotation from the page Wikiversity:Scope regarding teaching reads:

“The Wikiversity should become a premier source for accurate teaching materials — bibliographies, lesson plans, syllabi, effective animations, well written papers, ideas for in class and at home exercises, discussion topics, course work, etc. — as well as offer a clearing house for new and emerging teaching methodologies and technologies of learning. Teachers making use of Wikiversity materials are encouraged to comment on and adapt those materials in a continual effort to refine and improve Wikiversity’s resources.”

Researchers have reported experiences of developing and delivering a course through Wikiversity. The 10-week course Composing Free and Open Online Educational Resources communicated weekly
readings, quasi-formal assignments and feedback through Wikiversity. The course also used video conferencing and blogs and included some 70 participants from all over the world, though “towards the end of the course less than two dozen individuals were participating with any level of consistency”.

Wikiversity falls within the open education movement. Outside Wikimedia are other open education wikis, such as the WikiEducator (wikieducator.org). The ideas behind these efforts have been summed up in the Cape Town Open Education Declaration (www.capetowndeclaration.org).

Textbooks

Wikimedia Foundation project Wikibooks provides a site where editors may write free books collaboratively, e.g., open textbooks. Though a number of books have been declared ‘featured books’ the success has been mixed. Commentors have pointed to the difficulty in modularization of the text as compared to an encyclopedia as a reason for lack of major success.

A 2008 survey among 80 Wikibook contributors (based on 1’500 contacted) found that the respondents were mostly younger than 25 years (58%) and male (97.5%). Development of books are usually associated with high educational level, but the survey found that 39% only had a high school degree or lower. About one in four reported frustration with their wikibook project. 42% indicated that the book could not be completed.

Israeli researchers have reported on the development of an academic textbook on information systems using wiki technology. Faculty, undergraduate and graduate students participated during the course of two years using the MediaWiki software. The authors argue that the wiki textbooks leads to empowerment of both teachers and students over commercial publishers. About 1’200 students from 20 different classes and three different universities have used the text book. The wikitextbook was developed from an original outdated e-textbook consisting of 225 subsections that were converted to wikipages. The students read and augmented the pages and created new ones, so the number of pages would approximately double. The researchers found that for some classes the wiki contribution of students was associated with higher grades on average. They also found a long-tailed distribution of edit activity among the students. Interestingly, one of the successful Wikibooks, the one on the LaTeX markup language, was also seeded from an established work. Computer scientist Andrew Roberts had written an well-structured online tutoril called Getting to Grips with LaTeX available from http://www.andy-roberts.net/writing/latex. Wikibookians contacted Roberts, who gave permission to the use of his content on Wikibooks. This book is now one of the prime resources on LaTeX reaching top 5 in my search on Google and Bing with the query “latex”. So is the initial seeding the problem with Wikibooks? Would a good initial seed establish a modularization that Wikibooks contributors can use as a base for a successful work?

6 Open questions

What are the open questions in wiki and Wikipedia research?

The big equation?

Is it possible to setup an equation for when a wiki will evolve successfully? Or how successful it is? What variables are part of such an equation? Are money and gender part of the equation?

The Wilkinson-Huberman equation (see page 31) is surprising. Why would the number of edits on an ensemble of articles only depend on the age of the articles and a “random” fluctuation? What is this “random” fluctuation? Why are there no explicit external variables in the equation? Surely, an edit from a user may result in that edit showing up on the “my watchlist” of another user and possibly motivate that other user to make a new edit. In this case the edits are driven by a Wikipedia-intrinsic mechanism, but intuitively should edits not primarily be driven by notability of the subject? The Wilkinson-Huberman equation models the collective of an ensemble of articles of similar age, meaning that the model may have room for the notability of the subjects within the lognormal distribution. The modeling rested on Wikipedia data up to November 2006 in the period of Wikipedia large growth. Given the decline in the number of editors, does the model still hold? In the same work283,284 the two researcher show that quality articles are associated with more edits, so do quality articles attract more edits or do many edits results in a quality article?

Why are some Wikipedias bigger than others? The usual initial answer is that the size relates to the number of speakers of the language. This does not explain all variation, and Morten Rask introduced further features to explain the size variations: Human Development Index (HDI), Internet penetration and days in operations.125 These extensions can apparently not explain some size variations, e.g., the Norwegian bokmål Wikipedia had as of

52
January 2014 over 400,000 articles while the Danish well below 200,000 and the Swedish over 1.6 million. The differences in sizes of Nordic Wikipedias are mirrored in the size of Wikimedia Chapters: Wikimedia Norway has had almost 100 paying members of the organization, while the Danish chapter has had around 20 members. While differences in days of operation, language speakers, HDI and Internet penetration can to some extent explain the Wikipedia numbers it does not explain all. The bot activity with automated creation of articles on the Swedish Wikipedia explains part of its large size, but why is the Norwegian bigger than the Danish? Faroese Islands and Greenland share almost the same population, yet the Greenlandic Wikipedia shows very little activity with some 1,600 articles, while the Faroese Wikipedia shows more activity with over 7,000 articles. The activity of the Faroese Wikipedia seems to be related to the dedicated activity of a single individual.

What can explain the large difference in number of edits contributors make on Wikipedia? For employed knowledge workers, such as scientists and computer programmers, studies have reported individual variability in productivity with ratios in the tens,\textsuperscript{484–486} as well as reporting heavy-tailed distributions of productivity.\textsuperscript{484} Although these numbers are large, the individual differences in Wikipedia contributor output dwarf them with orders of magnitude. The usual Wikipedia studies focusing on contributor output do not measure productivity directly (as work per time unit), but rather the overall output, so the time used for Wikipedia contribution could possibly explain the huge differences. A report on Justin Knapp, the first Wikipedian to reach one million edits, said “he has spent 10 hours a day reading and adding to posts”.\textsuperscript{487} Is there a difference in productivity among Wikipedians? Studies on programmer productivity suggest experience, motivation and intelligence and “a general factor of programmer proficiency” as explanations for the large differences,\textsuperscript{485, 486} while William Shockley in his study on scientists considers an ‘organizational hypothesis’, mental combinatorial ability and a hurdle model with interacting mental factors.\textsuperscript{484} The organizational hypothesis (dismissed by Shockley for scientists), where heads of laboratories get coauthorships, is not relevant for Wikipedians as contributors cannot become coauthors on edits. Wikipedia’s suggestion for an “Attrition Pipeline” with “impediments to editing at each stage of editor lifecycle” would resemble the Shockley hurdle model, if there are individual differences in contributor persistence in overcoming each impediment, e.g., a struggle with wiki syntax or dealing with policy complexities. Other explanations for the observed variation build on the idea of ‘preferential attachment’, ‘success-breeds-success’ or ‘cumulative advantage’. If this explanation is the case then what is the mechanism of feedback? The experiment with barnstars (see page 29) produced only a relatively small variability in productivity between groups that did and did not receive a social reward. The Kaggle prediction competition Wikipedia’s Participation Challenge in 2011 sought “to understand what factors determine editing behavior”. In this competition participants should predict the future number of edits a contributor would make. One of successful competitors used only the edit history for prediction based on the features: number of edits in recent time periods, number of edited articles in recent time periods and time between first and last edit.\textsuperscript{488} Another system used over 40 features but did not yield better results.\textsuperscript{489} A deeper explanation for these results lacks: “You edit because you have edited” seems not a satisfactory explanation.

Can Wikipedia withstand contributions from professional public relations professionals and search engine optimization companies? Can the evolution of conflict-of-interests edits be quantified? In 2005 and again in 2006 Eric Goldman predicted that Wikipedia would fail within five years.\textsuperscript{182, 183} Yet again in 2010 he would point to the problems faced by the Open Directory Project as a user-generated content site becoming overwhelmed by relentless spam, providing a cautionary tale to Wikipedia.\textsuperscript{268} Seth Finkelstein would in 2008 write that he “sometimes remind people that ideological communes tend to end badly, too often with a few manipulative leaders extensively profiting at the expense of a mass of followers who lose everything”.\textsuperscript{224} Yet in 2015 Wikipedia still stands. Still in 2013 reporting from a major sockpuppert investigation journalist Simon Owen gave a pessimistic outlook: “But while reporting this article I couldn’t help comparing the sockpuppet discovery to a large drug bust—perhaps it might take out a major kingpin, but at the end of the day it’s a relatively minor victory in what is an otherwise losing war on drugs.”\textsuperscript{197} Furthermore, in the spring of 2015 Wikipedian and researcher Piotr Konieczny op-ed ed his growing concerns about “advertisements masquerading as articles” arguing that “Wikipedia has been gravitating towards a vehicle for business and product promotion for too long” and called out for a major cleanup drive of promotional articles which he with a napkin estimate put to 300,000 on the English Wikipedia, i.e., 3% of the total number of articles.\textsuperscript{490} The Sony email leaks from November 2014 cast some light
on companies Wikipedia practices, showing a Sony employee apparently updating Wikipedia articles on the CEO and the company encouraging creation of articles about their movies.\textsuperscript{491}

In 2013 Jimmy Wales pointed to what he saw as two threats for Wikipedia’s ongoing success: authoritarian governments censoring the Internet and “clumsy legislation driven by special interests”, such as the Stop Online Piracy Act.\textsuperscript{492} Are these threats serious?

What relation should there be between academics and Wikipedia?

What relation should there be between academics and Wikipedia? Should academics be encouraged to add their work on Wikipedia and rewarded or would a large-scale involvement of academics deteriorate Wikipedia because academics in search of employment and grants would tend to make biased conflict-of-interest contributions with overstating the importance of their own work? Does formal academic contributions short circuit the open collaborative hobby-oriented nature of Wikipedia editing? Should Wikipedia develop a method whereby experts can claim an article version as expert verified?

Since 2012 Wikiversity has incorporated a few peer-reviewed articles, e.g., “Diagram of the pathways of human steroidogenesis”, “Reference ranges for estradiol, progesterone, luteinizing hormone and follicle-stimulating hormone during the menstrual cycle” and the initial study “Average weight of a conventional teaspoon made of metal”\textsuperscript{493} all written by Mikael Häggström. On Wikipedia’s WikiProject Medicine talk page editors have discussed whether such articles can act as reliable sources for Wikipedia.\textsuperscript{494} The discussants would not fully embrace and recognize Wikiversity as reliable sources, — with its present form of peer-review. It is a question whether wiki-based publishing can gain traction and sufficient recognition, so that such a Wikiversity-based approach is a viable option for career-oriented scientists. Should such an approach be supported?

Why are not all scientific articles cited or mentioned on Wikipedia? Is it because the topics that the scientific articles cover are not yet dealt with in depth by Wikipedia, or is it because the topics of the majority of scientific articles are not notable enough and should be left out? Is much of science simply not encyclopedia-worthy? The Open Access Reader is a project initiated in 2014 aiming to systematically ensure coverage of ‘significant’ published open access research in Wikipedia.\textsuperscript{495} How feasible is such a project given the large number of scientific articles published each year?

Writing about scientific topics in Wikipedia one comes to question the validity of the claims in many scientific articles. Science itself has had discussions about the issue sometimes referred to as the ‘reproducibility crisis’ in experimental sciences.\textsuperscript{496} Ioannidis claims from theoretical considerations and a number of assumptions that “most published research findings are false”,\textsuperscript{497} — ironically a claim itself criticized as false.\textsuperscript{498} However, he has done empirical research finding that a considerable part of highly-cited biomedical does not replicate in subsequent studies.\textsuperscript{499} Wikipedia already has a warning in the content guideline Wikipedia:Identifying reliable sources against relying on individual studies: “Isolated studies are usually considered tentative and may change in the light of further academic research.”\textsuperscript{496} Is Wikipedia more accurate than the average scientific article? If so, can the Wikipedia process be transferred to the scientific process? Can Wikipedia and similar tools do anything about the ‘reproducibility crisis’ in experimental sciences?

Although citations and universities may be ranked to some degree based on citations in links in Wikipedia,\textsuperscript{500,148} scientists have poor coverage on Wikipedia.\textsuperscript{69} Should we caution against using Wikipedia for altmetrics? And if Wikipedia altmetrics becomes established would that have a deteriorating effect of the quality of Wikipedia? How do we ensure that Wikipedia continuously reflects the knowledge in the academic literature?

Small scientific wikis, independent on large organizations such as the Wikimedia Foundation, may run into sustainability problems like other online scientific databases; The wikis are created and operated not solely by research groups on time-limited research funding, and if the funding stops the continued operation of the wiki is jeopardized. Open wikis may require daily attention which is not available from researchers moved on to new projects. A few scientific wikis have stopped operation, e.g., in January PDBWiki produced the message “PDBWiki is offline as of January 14th, 2014\textsuperscript{151} Also WikiProtein seems no longer available. It is an open question whether Wikipedia and its sister projects, particularly now with Wikidata, can provide a better framework for sustainable scientific wiki work compared to separate wikis. Surely, if applying Wikidata instead of a locally

\textsuperscript{491}“Nomination” of steroidogenesis article in Wikiversity to be used as reference section on “Wikipedia talkWikiProject Medicine”; 22:58, 15 April 2014, oldid 604565183.

\textsuperscript{494}According to [\url{http://pdbwiki.org/}]

\textsuperscript{495}Wikipedia:Identifying reliable sources, oldid = 596070076, English Wikipedia.

\textsuperscript{496}https://meta.wikimedia.org/wiki/OpenAccessReader.
controlled wiki the scientist’s control of authorship is mostly lost but the task of system and wiki administration is considerably reduced. It is unclear whether Wikidata sufficiently supports the many varied data representations found in scientific wikis and whether external tools can sufficiently present the Wikidata data in the ways its presently done on specialized scientific wikis. Furthermore, will notability issues hinder the representation of scientific data on Wikimedia sites?

What can Wikipedia say about the world and can it affect the world?

To what extent can Wikipedia be used for for prediction of world events such as movie box office, election results, etc.? Does Wikipedia have a positive societal value (whatever that is) or is it merely a Friday bar argument settling site? How appropriate is the two digit—or even four digit—value in milliard US Dollars mentioned by Band and Gerafi for the consumer value or positive externalities?

Does Wikipedia affect the world and to what extent? The Chinese censorship of the Chinese Wikipedia and the Iranian censorship of some of the Persian Wikipedia articles could indicate that some authorities are afraid of the impact of Wikipedia information. To what extent does political Wikipedia articles affect voting?

Noam Cohen attributed the Microsoft Encarta demise to the competition from the free Wikipedia. The introduction of Encarta had earlier — in the 1990s — impacted the sale of the competing Encyclopædia Britannica: Encyclopædia Britannica hard copy book sales fell by 50% over a couple of years, affecting an annual revenue on several hundreds of millions US Dollars.

Do students get better by using Wikipedia? Researchers have carried out a field study with a controlled experiment using OLPC computers, that among the applications had an offline copy of Wikipedia. Among several hundreds Ethiopian children the researchers found that children with no laptop got higher grades in mathematics. They also found that school engagement was no different among the two groups with and without an OLPC device. Could this indicate that children are put at a disadvantage by reading Wikipedia? This is not what the study says, because the researchers also examined abstract reasoning in the children and reported that the results indicated that children with laptops outperformed children without. Other technologies have been promoted in learning, but although heralded as a new age for learning, they may not fulfill the promises to revolutionize education. A study on iPad reading in Year 6 pupils concluded “[a]lthough the students participating in the study reported an increase in engagement when using the iPad, there was not a corresponding rise in achievement.”

Comparisons of the scores from the Programme for International Student Assessment (PISA) in Denmark with pupils grouped according to whether they had access to tablet computers in school or not showed a lower average score for those with tablet access. Such studies call into question whether Wikipedia furthers education, and whether teachers should encourage students to use Wikipedia at all. Given that at least the spikes in Wikipedia page views are related to pop culture phenomenons, such as celebrity deaths and Super Bowl halftime entertainers are students diverged away from learning ‘useful knowledge’ to reading pop culture when they engage with Wikipedia?

Structured data?

In January 2013 the Wikidata project was launched and it grew quickly. In terms of pages it quickly surpassed the English Wikipedia. A few papers have described the system in overview, but otherwise very little research have been done on this young project. What is the authorship and readership of Wikidata? Are bots the dominating force on Wikidata or do humans have a say? What types of data can humans contribute with? Do we see a highly skewed editing pattern where a few prolific bots make the majority of the work?

Will we see ‘ontology wars’ on Wikidata where editors cannot agree on properties? One example of property discussions is how humans/persons should be described. Initially, the property ‘GND-type’ (German National Library) stating ‘person’ was used but later this property was deprecated for labeling persons as persons.

What about the multi-lingual nature of the Wikidata project? Is the regimensal description of items across languages “dangerous”? Does an ontology uniform across languages prohibits cultural diversity? Is the predominantly English discussions on Wikidata a problem for non-English users? Before the launch of Wikidata Mark Graham raised concerns that the “highly significant and hugely important” changes brought by Wikidata “have worrying connotations for the diversity of knowledge” on Wikipedia. He believed that “[i]t is important that different communities are able to create and reproduce different truths and worldviews,” exemplifying the problem with the population of Israel: Should it include occupied and contested territories? Wikidata—at least partly—counters Graham’s worry: each property (e.g., ‘population’)
may have multiple values, and qualifiers associated with Wikidata values can distinguish between the scope of each claim. Denny Vrandečić, project director of Wikidata, explained: “We do not expect the editors to agree on the population of Israel, but we do expect them to agree on what specific sources claim about the population of Israel.” However, the question remains whether the Wikidata system provides sufficient flexibility to capture, e.g., the slight interlanguage differences in some concepts, and when there are differences does the definition fall back to one centered in the anglocentric world-view? Take the Germanic concept of ‘Hochschule’/’hejskole’/’högskola’. The English Wikipedia has a separate article about the German(ic) concept of ‘Hochschule’, and Wikidata has an item for the concept. However, that concept dangles in Wikidata, since the German Wikipedia ‘Hochschule’ article leads to another Wikidata item associated with the concept of a ‘higher education organization’. Another potential problem could be whether certain values for a property should exist or not. Take the borders of Israel as an example: The property ‘shares border with’ presently lists Syria, Jordan, Egypt and Lebanon, and not the State of Palestine. The international recognition of the State of Palestine as a state varies between countries, so according to one view the State of Palestine should be listed, while the opposing view would hold it should not. A suitable qualifier (‘as recognized by’?) could possibly resolve it. Another issue arises for almost similar concepts which can be linked by the informal interwiki Wikipedia links with no problem, but where the semantic description will be difficult in a multilingual environment. Take the classic scientific article *The magical number seven plus or minus two: some limits on our capacity for processing information*. In the English and French Wikipedias it has its own article, while the German Wikipedia chooses to make an article on the psychological concept described in the article (“Millersche Zahl”): Whether this item has an author or not and a date of publication depends on whether one regard it as a publication or a concept. A pure semantic approach would split such cases, increasing babelization.

I have experienced a few instances of vandalism on Wikidata: French footballer Anthony Martail called ‘Rien’ (French: nothing, actually a Norwegian lake) and American basketball player Michael Jordan called ‘insect’. ‘Called’ here means attributed the GND type of ‘Rien’ or ‘insect’, — neither standard GND types. Will vandalism be a problem on Wikidata? On Wikipedia a bot with machine learning-based detection from vandalism features operate. Would automatic vandalism detection be possible on Wikidata?

Can Wikidata describe everything? What kinds of data can Wikidata not conveniently describe? Wikisource, Wikivoyage and Wikimedia Commons got relatively unhindered their language links represented in Wikidata and according to the Wikidata development plan for 2014+ Wikiquote, Wikinews, Wikibooks and Wikiversity are schedule for inclusion in Wikidata. However, Wiktionary was as of January 2014 not scheduled for Wikidata inclusion. It is unclear if the item/property system of Wikidata is an appropriate representation for words and how lexemes, lemmas, forms and senses should most easily be represented with the Wikidata data model.

How advanced queries can be made with Wikidata data? Magnus Manske’s *AutoList* tool can already now carry out on-the-fly queries like “all poets who lived in 1982”. But can such queries continue to be carried out effectively, and will Wikidata generally be able to scale? For example, how will Wikidata cope with several thousand claims per item? It may be worth to remember that Wikipedia articles seldomly reach past 200–300 kB because articles get split into subarticles, e.g., “Barack Obama” splits into “Family of Barack Obama” and “Illinois Senate career of Barack Obama” etc. This sharding technique seems not to be readily possible with Wikidata. Dynamic Wikidata-based translation in its present form, e.g., through the *qLabel* Javascript library, can result in multiple requests to Wikidata servers for just a single page view on a third-party website. If successful, will Wikidata-based translation results in unmanageable load on Wikidata?

**Acknowledgment**

Thanks to Daniel Kinzler, Torsten Zesch, Felipe Ortega, Piotr Konieczny, Claudia Kołtzenburg and James Heilman for pointing to references and tools. Thanks also to Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari and Arto Lanamäki with whom I am writing systematic reviews about Wikipedia research.¹⁴

**References**

ciety for Scientometrics and Informetrics : 10th, 2005.


Annotation: Description of a system that extracts information from the templates in Wikipedia, processes and presents them in various ways. Some of the methods and services they use are MySQL, Virtuoso, OpenCyc, GeoNames, Freebase, SPARQL and SNORQL. The system is available from http://DBpedia.org


Annotation: Reports the result of a survey on Wikipedia use among American adults.


Annotation: Presents the results of a survey on Americans use of Wikipedia. It is based on telephone interviews conducted in the spring 2010. It is found that 53% of Internet users used Wikipedia, corresponding to 42% of adult Americans.


Annotation: Short blogpost plotting the number of theses form major Chinese-speaking regions as a function of year.


Annotation: Short article that sets up a framework for systematic review of Wikipedia research identifying roughly 1,000 articles.


Annotation: Short review of peer-review Wikipedia research. The researchers identified over 400 academic papers. A few of these 400 are briefly summarized.


Annotation: A review of Wikipedia research.


Annotation: Systematic review on research on Wikipedia focusing on peer-reviewed journal articles and theses published up until June 2011 and included some important conference papers. The review covers all aspects of Wikipedia research, including uses of Wikipedia-derived data.


**Annotation:** Systematic review on scientific research on Wikipedia and its readership.


**Annotation:** Data set with page view statistics from the English Wikipedia collected in March and April 2015.


**Annotation:** Describes a dataset with Wikipedia text represented in XML.


**Annotation:** Discuss several aspects of history on the English Wikipedia and how professional historians should regard that wiki. The author also make a quality assessment of a American history articles on Wikipedia and compare them against Encarta and American National Biography Online.


**Annotation:** Blog post on Wikipedia page view statistics discovered to be wrong.


**Annotation:** Blog post that 'explores the complications of attribution and identification in online research'.


**Annotation:** Book on various aspect of Wikipedia and related phenomena: Usenet, Nupedia, wikis, bots, Seigenthaler incident, Essjay controversy, Microsoft Encarta, etc.


**Annotation:** Book about MediaWiki and Semantic MediaWiki.


**Annotation:** Aggregation of the Wikimedia Research Newsletter for the year 2012 written by various researchers.


**Annotation:** Describe the German Wikipedia based on qualitative interviews with seven members of the Berlin Wikipedia community.


**Annotation:** Discuss commons-based peer production exemplified with free and open source software, SETI@home, NASA Clickworkers, Wikipedia and Slashdot.


[40] Andrew Lih. How Wikipedia solved the knowledge gap. TEDx-American University, April 2014.

**Annotation:** Talk on Wikipedia at TEDx-American University.


**Annotation:** An analysis with respect to information quality of a sample of the discussion pages on Wikipedia. The analysis is based on the their own information quality assessment model, and they provide example quotations from the pages. They conclude that ‘the Wikipedia community takes issues of quality very seriously’.


**Annotation:** Report of a comparison of the accuracy in Wikipedia and Encyclopedia Britannica


**Annotation:** Report on three experts informally reviewing popular culture, current affairs and science articles in Wikipedia. Popular culture reviewer calls it “the people’s encyclopedia”; current affairs reviewer “was pleased by Wikipedia’s objective presentation of controversial subjects”, but cautioned “a healthy degree of skepticism and skill at winnowing fact from opinion is required”. Lastly, the science reviewer found it difficult to characterize because of the variation, noting “flaws” and “the writing is not exceptional”, but “good content abounds”.


**Annotation:** Summary of the research by Lara Devgan et al. “Wiki-Surgery? Internal Validity of Wikipedia as a Medical and Surgical Reference”.


**Annotation:** Examines Wikipedia as a drug reference with a comparison against Medscape Drug Reference. Wikipedia has more omissions but no factual errors among 80 drug-related questions/answers, e.g., for administration, contraindications and issues around pregnancy/lactation. They also find that Wikipedia has no dosage information, but this is not surprising given that Wikipedia explicit encourage authors not to add this information. Four factual errors were found in Medscape. Two were conflicting information in different...
parts of the text while the remaining two were due to lack of timely update.


Annotation: A small blinded comparison of 3 Wikipedia articles for medical student information against AccessMedicine, eMedicine and UpToDate online resources. Wikipedia was found to be unsuitable for medical students.


Annotation: The same study as “Putting Wikipedia to the test: a case study”.


Annotation: A short abstract reporting that Wikipedia had similar accuracy and depth compared to the professionally edited information in the National Cancer Institute’s Physician Data Query (PDQ), but Wikipedia was less readable as evaluated with the Flesch–Kincaid readability test.


Annotation: An analysis of the quality of Wikipedia articles in the nephrology area. They looked at comprehensiveness, reliability and readability.


Annotation: Study on the quality of information on Wikipedia about mental disorders in terms of accuracy, up-to-dateness, coverage, references and readability with comparison against 13 other websites. 10 topics were rated the three psychologists with relevant expertise. Wikipedia was generally rated higher than the other websites.


Annotation: News articles about an expert evaluation of 134 articles on the Finnish Wikipedia on the dimensions lack of errors, coverage and balance, sourcing, topicality, neutrality and clarity.


Annotation: Study on the quality of Wikipedia articles in the nephrology area. They looked at comprehensiveness, reliability and readability.


Annotation: A study with comparison of assertions in Wikimedia articles on medical conditions with assertions in the peer-reviewed literature.

Annotation: Editorial about study on Wikipedia quality for medicinal conditions.


Annotation: Analysis of 22 Wikipedia articles on prescription drugs and whether they are updated in relation to new drug-safety communications from the U.S. Food and Drug Administration. 36% articles remained unchanged after more than a year.


Annotation: Analysis of the quality of drug information in the German and English Wikipedias. Accuracy, completeness, readability, number of references, revisions and editors were considered.


Annotation: A study where 258 researcher were asked to evaluate a Wikipedia article. 69 researcher responded and evaluated either an article within their own field of research or a random article. The evaluation focused on credibility.


Annotation: Danish newspaper article comparing eight articles in the Danish Wikipedia with eight articles in Den Store Danske encyclopedia, with Wikipedia performing slightly better.


Annotation: Discussion of electronic reference works.


Annotation: A comparative study of German Wiktionary, OpenThesaurus and GermaNet.


Annotation: Statistics on the outbound scientific citation from Wikipedia with good correlation to the Journal Citation Reports from Thomson Scientific.


Annotation: Study of the citations in Wikipedia to scientific journals. Wikipedia articles and journals are bi-clustered and a longitudinal analysis of the number of citations is performed.


Annotation: Examination of the coverage of Wikipedia by two methods: First method entails sampling 3000 Wikipedia articles and manually categorizing them against the Library of Congress category system, then comparing them with the number of published books within the categories. This method shows that, e.g., the literature category is underrepresented in Wikipedia, explained by the lack of fiction in Wikipedia. Other categories underrepresented are, e.g., medicine, philosophy and law, whereas science and some areas with systematic mass-inserted articles, such as naval and geography, are well represented. Also fans tend to add, e.g., popular music groups and literature description articles. The length of medicine and law article are large compared to the rest of the articles. The second method matched Wikipedia articles with specialized encyclopedia in physics, linguistic and poetry and found missing articles in Wikipedia.


Annotation: A comparison of Stanford Encyclopedia of Philosophy, Internet Encyclopedia of Philosophy and Wikipedia of twentieth century philosophers. Wikipedia had by far the most philosophers, and Wikipedia had also a large number of more recently born philosophers. Besides this the three encyclopedia did not differ greatly in coverage with respect to gender, disciplines and nationality, though Wikipedia had a smaller fraction of German and French but a higher fraction of other Europeans, and Wikipedia has also categorizes some in the spiritual disciplines as philosophers.


Annotation: Study on the comprehensiveness of Wikipedia in ICD-9 and ICD-10 diagnostic codes in gastroenterology. Over 80% were covered. Readability is also considered.


Annotation: Describes a system for finding articles on Wikipedia that maps to parts in the LOINC database.


Annotation: Study on the possible gender bias in the the content of Wikipedia with examination of several thousand biographic articles in comparison to Britannica and other resources.

Annotation: Analysis of the coverage of scientists on Wikipedia. Wikipedia presence is compared with data from Scopus.


Annotation: Show news article with interview of Jimmy Wales.


Annotation: A study with 135 medical students answering clinical questions with the help of Web resource. Google and other search engine were most effective. Sites such as Wikipedia, eMedicine and MDConsult less, though they sometimes provided the ultimate answer, and Wikipedia did that the most.


Annotation: Examines the length of sets of Wikipedia articles and compares them with a number of other variables: With year, Encyclopedia Britannica, country population and company revenue.


Annotation: A study on a range of quality of scientific articles on the English Wikipedia along a number of dimensions, e.g., coverage, referencing, length, user perception.


Annotation: Describes an algorithm for determine the topic distribution of a Wikipedia article over top-level categories based on the Wikipedia category graph, and evaluate it against human labeled data. By applying the algorithm to the 2006 and 2008 data sets of the English Wikipedia the topic coverage over time is described, and with combining the method with a degree of conflict page labeling the most contentious topics are found.


Annotation: Analysis of how the gender gap of Wikipedia contributors was represented and discussed in media.


Annotation: Small report on the proportion of males in Wikipedia biographies about living people.


Annotation: Report of an analysis of the gender ratio on Wikipedias by using Wikidata material.


Annotation: Blog post reporting a study with Piotr Konieczny on gender representation in Wikipedia.


Annotation: Blog post with an analysis of the representation of gender in Wikidata.


Annotation: Blog post with a Wikidata-based analysis of gender bias in Wikipedias.

[104] Han-Teng Liao. preliminary results from the Wikipedia Gender Inequality Index project - comments welcome. Wiki-research-l mailing list, January 2015.
Annotation:


Annotation: Estimates the number of species.


Annotation: Press release stating that the chemical compound database has reached its 60 millionth entry.


Annotation: Describes the quality of history articles on the Danish Wikipedia. A university teacher monitored articles created by his students as they were edited by other Wikipedians.


Annotation: Short comment on Wikipedia publishing news before it becomes published by general media.


Annotation: Describes by a few examples how Wikipedia is used as a source in news media and other domains. The article notes that editor at Philadelphia Inquirer warned journalist never to use Wikipedia "to verify facts or to augment information in a story". It mentions the WikiScanner and the case of Seigenthaler. A number of different people in the news business are interviewed expressing different opinion on the use of Wikipedia in news media, and studies by Nature and Roy Rosenzweig are described briefly.


Annotation: A study examining the inbound citations to Wikipedia from press articles published in 2003 and parts of 2004. It displays "diversity" (number of individual editors of an article) against "rigor" (number of edits for an article) in some Wikipedia articles and the study regard these as indicators of quality. World War II, Islam and Astronomy scored high on these dimensions. The indicators of article quality increased after press citation for some Wikipedia articles.


Annotation: Reports on a genre analysis of Wikipedia compared with Everything2 and Columbia Encyclopedia. A quantitative analysis of the formality of the text is performed by counting words related to formality or informality. They find that the style of Wikipedia articles is close to that of Columbia Encyclopedia.


Annotation: A non-quantitative, descriptive study on a few Wikipedia articles in terms of structural form (i.e., genre) and its evolution as the articles are extended.


Annotation: A quantitative test of conformity with the W3C’s Web Content Accessibility Guidelines of 100 Wikipedia articles and 265 articles on the Web referenced by Wikipedia. The 100 Wikipedia articles had on average a better score of accessibility.


A case in the Supreme Court of India where Wikipedia is used for the definition of ‘laptop’.


Annotation: Describes Wikipedia as a research tool for students and lawyers.


Annotation: Danish article with an interview of Jimmy Wales as the founder of Wikipedia.

Morten Rask. The reach and richness of Wikipedia: Is wikinomics only for rich countries. First Monday, 13(6), June 2008.


Annotation: Interview with Executive Director of the Wikimedia Foundation Sue Gardner.


Annotation: A news article on the issue of why the Arabic Wikipedia is relatively small.

Xan Rice. Internet: Last piece of fibre-optic jigsaw falls into place as cable links east Africa to grid. guardian.co.uk, August 2008.

Annotation: News article about a new submarine cable for communication for the poorly connected east Africa.


Annotation: Blog article reporting about the work of a computer program, a bot, creating many articles on species on the Swedish Wikipedia.


Annotation: Describes so-called ‘small-world’ networks which can be constructed as interpolations between regular (lattice) and random networks. They exhibits short averages paths while still being clustered. Examples are given by film actors, power grid and the neural network of a worm together with dynamic networks, e.g, multiplayer prisoner’s dilemma.


Annotation: Network analysis of the coauthorship network in English Wikipedia, where the main contributors to a Wikipedia article is identified with a method of Adler et al. and various network metrics are extracted across time and topics.


Annotation: Network analysis of intrawiki links on a number of different language versions of Wikipedia and report a range of network characteristics.


Annotation: Description of the use of the Wikipedia category graph for determining semantic relatedness. Several different similarity and distance measures on the graph are examined on several human labeled datasets from the German Wikipedia. Graph characteristics (average shortest path, cluster coefficient and power law exponent) are also shown.


Annotation: Network analysis of the user interaction on the article and user talk pages of the English Wikipedia.


Annotation: Network analysis of the intrawiki links on the English Wikipedia.


Annotation: Analysis of various aspects of early Wikipedia, especially the number of articles and categories.


Annotation: Discuss accelerating networks with network analysis of the number of nodes and links, exemplifying it on three different language versions of Wikipedia.


Annotation: A multivariate analysis of the wikilink graph of the English 2005 Wikipedia with around 500'000 pages with a discrete version of the hubs and authority algorithm.


Annotation: Examines the PageRank and the ‘reverse’ PageRank (CheiRank) for links within Wikipedia as well as a combined rank they call 2DRank.


Annotation: Analysis of intrawiki links on Wikipedia across several years.


Annotation: Network analysis of the bibliographic pages across the 15 largest Wikipedias.

Annotation: Network analysis and visualization of the link network among Western philosophers described on Wikipedia.


Annotation: Describes algorithms for large-scale eigen decomposition and applies them on a large document-term matrix constructed from the English Wikipedia.


Annotation: Brief article describing topic mining with latent Dirichlet allocation using LingPipe on Wikipedia text as available in the Wikipedia for Schools dataset.


Annotation: Visualization of relationships between Wikipedia pages based on co-authorship patterns.


Annotation: Slides about neuroinformatics and data clustering of Wikipedia content.


Annotation: Study with an application of an algorithm for identification of maximal bicliques on the network of articles and editors on the English Wikipedia.


Annotation: Analysis of geotagged Wikipedia articles about topics in the with Middle East and North Africa (MENA) region and established the co-editing social network. By examining the user page they tried to determine if the user was from a specific country. They found that most editors were from the West rather than MENA countries.


Annotation: Describes a social network analysis by looking at coauthor graph in Wikipedia from sequential editing.


Annotation: Short description of a network visualization of edits in Wikipedia including sentiment analysis of the edits.


Annotation: Description of an online service with sentiment analysis of Wikipedia edits for monitoring business-related Wikipedia articles and display of the results with network visualizations.


Annotation: Report on a study that injected false fact in Wikipedia. Of 36 errors only 15 was removed within 48 hours.


Brian Brady. BBC staff rewrote Wikipedia pages to water down criticism. The Independent, December 2007.


Annotation: Description of a opinion mining system for Wikipedia edits.


Annotation: Reports on a survey among 1'284 public relations/communications professionals about their perceptions and practice with Wikipedia.


Annotation: News article on Wikipedia subcategorizing female but not male American Novelists.


Annotation: Description of the WikiDashboard web-site that makes a visualization of author contributions of a Wikipedia article and embeds it in a proxy copy of the Wikipedia article. Social transparency in relation to Wikipedia is also discussed.


Annotation: Describes a model for time-dependent user reputation on wikis and applies it on the entire revision history of the English Wikipedia.


Annotation: Computation of the Gini coefficient for author contributions for the top ten largest Wikipedias. They find the Gini coefficients to be from 0.92 to 0.97.


Annotation: A report on a small investigation on the different amount of contribution authors make to Wikipedia. Contrary to previous thought the examination of a few Wikipedia articles shows that essential content was mostly written by occasional contributors.


Annotation: A study on the amount of edits on the English Wikipedia in 2009 made by bots or through assisted editing tools. Combined they make up 28.49% of the edits, — a number slightly larger than the number of edits performed by anonymous users.


Annotation: Describes experience with the CoWeb wiki used for teaching English. Students using the wiki had better attitude towards collaboration and got higher grades compared to another online environment.

Annotation: Describes experience with having freshmen-level college students write and interact in a wiki where the students can comment of each others articles.


Annotation: Report on a survey among several thousand Wikipedia editors.


Annotation: Report on an interview study on 32 Wikipedians contributing to Wikipedia health-related articles.


Annotation: Reports on the gender gap among Wikipedia contributors and discusses its influence on the topic emphasis on Wikipedia.


Annotation: Reports on a survey of FLOSS developers.


Annotation: Application of the Big Five Inventory and Real-Me personality questionnaires to 139 Wikipedia and non-Wikipedia users. The recruitment was based on targeting posting of links. Wikipedians scored lower on agreeableness and higher on openness. Differences in extroversion and conscientiousness depended on the sex of the subject.


Annotation: Analysis of the geotagged articles in 44 different language versions of Wikipedia.


Annotation: Discuss some aspects of digital media and higher educutions.


[229] Paolo Magrassi. Free and open-source software is not an emerging property but rather the result of studied design. arXiv, November 2010.

Annotation: Discusses free and open-source software and argues for a role of top-down coordination.


An analysis of 140 Free and Open Source Software project hosted on SourceForge with a study of contributor interaction network.


Annotation: Book about mass collaboration among animals and humans.


Annotation: Discussion of history of reference work production.


Annotation: Discuss the idea of stigmergy in relation to Wikipedia.


Annotation: Statistics on the worlds most visited Web sites during September 2006.


Annotation: A study of 35,000 active Arabic language blogs.


Annotation: Describes an investigation of search engine ranking for health topics in Wikipedia with search engine optimization methods. With queries selected from MedlinePlus, NHS and NORD they show that the English Wikipedia is more often on the first place for the selected queries on Google as compared to .gov domain, MedlinePlus, Medscape, NHS Direct Online and a number of other domains. The also investigated health-related topics with seasonal effects.


Annotation: Content analysis of Wikipedia articles on companies as well as examination of Internet search engine ranking. 10 Fortune 500 companies were selected in the years 2006, 2008 and 2010 and examined for topic, tonality and length, number of edits, etc.

[244] Nathan Safran. Wikipedia is the SERPs: Appears on page 1 for 60% of information, 34% transactional queries. Conductor Blog, March 2012.

Annotation: Reports a non-peer-reviewed study on the the position of Wikipedia in Internet search engine results. Using 2,000 keywords from representative search engine queries it was found that the appearance of Wikipedia on page one depended on whether the query was information or transactional and on how many words were in the query.


Annotation: Non-peer-reviewed study on the search engine ranking of Wikipedia. Wikipedia appear on Google’s page one based on 1,000 random nouns queried.

Annotation: ASCI benchmarks from 2010 to 2013 for websites such as Wikipedia, Pinterest, YouTube and Myspace.


Annotation: News report on a survey conducted by the Mannov company on the use of social media in professional contexts of Danish medical doctors.


Annotation: Report of Danish medical doctors use of social media for professional activities.


Annotation: Analysis of Wikipedia page view statistics.


Annotation: Review of studies on the economic value of Wikipedia and some further estimates on replacement cost and consumer value.


Annotation: Studies gender difference in the use of Wikipedia.


Annotation: Blog post on paid editing on Wikipedia by a Wikimedia Foundation employee.


Annotation: Overview of freelance writer’s word, page and project rates for different kinds of work.


Annotation: Describes interviews with 9 Wikipedia contributors and some of their characteristics: Most of the contributors tell that their initial edit was for correcting a problem or extending a weak article. As novices they were not aware of the Wikipedia community. As more experienced contributors they get a sense of community and decrease article writing and increase administration.


Annotation: Semi-structured qualitative interview with six Danish Wikipedia volunteers for investigating the motivation for contribution. The interviews were transcribed ‘categorized and analyzed thematically’.


Annotation: Reports on a survey among Wikipedia contributors about their motivation for sharing knowledge.


Annotation: An experiment where Wikipedia editors were given informal awards to see how it affected their productivity.

[263] Haiyi Zhu, Amy Zhang, Jiping He, Robert E. Kraut, and Aniket Kittur. Effects of peer feedback on contribution: a field experiment in
Annotations:


Annotation: Early notes about the VisualEditor project for a WYSIWYG editor in the MediaWiki software.


Annotation: News report on the controversial Wikipedia editing of the company Wiki-PR.


Annotation: From interviews of 22 Wikipedia contributors to understand their motivation the study argues that the incentive system in Wikipedia resembles that of science.


Annotation: Statistics on Linux Kernel development.


Annotation: Discussion of (possible future) sustainability problems for Sustainability and what can be done to counter them.


Annotation: Reports on an analysis of a survey among Wikipedia contributors for investigation of their motivation as contributors.


Annotation: A suggestion for a computation of reputation of the individual authors in a wiki-system. Formulas are shown but it is not clear if the suggestion is tested on a real wiki.


Annotation: Propose to view online participation under the framework of gateway theory as in contrast to antecedents-based explanations.


Annotation: Blogpost which reflects on the Essjay controversy in Wikipedia.


Annotation: From interviews of 22 Wikipedia contributors to understand their motivation the study argues that the incentive system in Wikipedia resembles that of science.


Annotation: Analysis of the effect of social awards on productivity in Amazon Mechanical Turk, the programming language Scratch and in Wikipedia.


Annotation: Description of various instances of Internet addictions.

Annotation: A quantitative longitudinal study of Wikipedia users.


Annotation: Results about a survey on Wikipedia contributors that have stopped editing.


Annotation: Data analysis of the evolution of Wikipedia.


Annotation: An investigation of what signals quality articles on the English Wikipedia. A growth model is suggested for the number of edits on an article, and this lognormal model shows good agreement with the observed number of edits. It is also shown that older articles typically have more edits than newer articles and that quality articles have comparably more edits, many distinct editors, many edits in its talk page, a quick ‘turnaround’ and a high number of edits per editor when correcting for article age and Google PageRank. They write that featured articles with high Google PageRank have a lower number of talk page edits than articles with low PageRank. It is not clear if that effect is due to archiving.


Annotation: Discusses the decline of USENET


Annotation: A description of research on why Wikipedia was successful and other crowd-source online encyclopedias were not. The encyclopedia as a familiar product, focus on substantive content rather than technology as well as low transaction cost in participation were reported.


Annotation: Discussion of reasons for success in Wikipedia, Slashdot and SPIP focussing on low transaction costs, clear vision and mechanisms and software tools to counter hostile contribution.


Annotation: Related to the paper A quantitative study for online encyclopedias: comparison of Wikipedia and Knol. The present paper has information, e.g., about which articles the researchers studied.


Annotation: Comparison of 20 articles from Wikipedia and Google Knol with respect to page views, number of words, readability, number of citations, number of references and types of references.

Annotation: News articles about Google announcing the closing of Knol.


Annotation: A brief note reporting that artificial markets, such as Hollywood Stock Exchange and Foresight Exchange, are good predictors of future outcomes.


Annotation: Discusses Linus’ Law with respect to Wikipedia.


Annotation: Text sentiment analysis of blogs by using Wiktionary, verb classes and support vector machines classifier.


Annotation: Conference article of ‘Extracting content holes by comparing community-type content with Wikipedia’.


Annotation: Describes a system for discovery of missing subtopics in a text by comparison against text in Wikipedia.


Annotation: Short paper on the same topic as ‘Extracting content holes by comparing community-type content with Wikipedia’.


Annotation: Describes the Koru document retrieval system, that is based on a thesaurus constructed from Wikipedia data and features automatic query expansion. A user study with a comparison system shows that Koru performs better and more preferred.


Annotation: Describes an information retrieval systems with sentiment analysis and query expansion using Wikipedia.

Annotation: Description of a system for combined text and image similarity computation evaluated on the ImageCLEFphoto 2008 data set and demonstrated on Wikipedia data.


Annotation: Describes the multilingual image retrieval system WikiWord based on data from Wikipedia.


Annotation: Description of the Synarcher program that analyze Wikipedia with the Kleinberg HITS algorithm and present related terms to a query term in a graph visualization.


Annotation: Describes an online named entity extraction system with disambiguation using YAGO2, Stanford NER Tagger and PostgreSQL.


Annotation: Describes ANGELINA3 that does notable person detection by looking up a person on Wikipedia, and the sentiment about the person is gauged by looking the person up on Twitter and using the AFINN word list for text sentiment analysis.


Annotation: Short description of a machine translation system for translating Swedish Wikipedia to Danish.


Annotation: News article explaining the at that time recently introduced term ‘folksonomy’.


Annotation: Analysis of the Wikipedia category system.

**Annotation:** Describes a semi-automated ontology learning system based on Wikipedia data.


**Annotation:** News articles noting large activity in US vice presidential candidate Sarah Palin Wikipedia article just before her candidacy became public knowledge.


**Annotation:** News story about an edit on Wikipedia that seemingly revealed information before it became public knowledge.


**Annotation:** Describes a realtime breaking news monitoring service using Wikipedia edits.


**Annotation:** Analysis of social media, (YouTube, Facebook, Twitter and Wikipedia), in relation to social media engagement in health care.


**Annotation:** Blog post with prediction of 2014 winners for 5 Academy Awards categories.


**Annotation:** Analysis of how good Google search volume and Wikipedia page views are for the prediction of elections in Iran, Germany and United Kingdom.


**Annotation:** Altmetrics study of Wikipedia extracting scientific journal article citations based on PMID and DOI with comparison to Faculty of 1000 data. With longitudinal analysis they found articles to the cited increasingly faster.


**Annotation:** Describes a system from sharing mutable data.


**Annotation:** Sets up a statistical model (a dynamic Bayesian network) for modeling of Wikipedia article trust by looking at the revision history. The set Bayesian priors based on whether the user is administrator, registered user, anonymous user or blocked user. Further the trust is based on the amount of insertion and deletions.


**Annotation:** Suggests word count as a simple measure for quality and shows that it perform well when classifying featured and random articles, — seemingly better than the methods of Zeng and Stvilia.


**Annotation:** A logistic regression analysis of which features were important for quality article on Wikipedia.


**Annotation:** News articles about bots operating on Wikipedia.


**Annotation:** Describes a reputation system for Wikipedia which enable individual words to be colored according to computed ‘trust’.


Annotation: Report from a prediction competition on Wikipedia vandalism detection. The corpus was based on both English, German and Spanish Wikipedias. Three systems participated.


Annotation: Description of the creation and addition of over 8000 gene articles in Wikipedia with an automated bot. Information is aggregated from Entrez Gene and a gene atlas for the mouse and human protein-encoding transcriptomes.


Annotation: Describes the WikiProject RNA where data and annotation about RNA are shared between Wikipedia and the Rfam database.


Annotation: Shortly describes shortly the WikiOpener MediaWiki extension, that can be used to include data from web-based database in a wiki. They state that the extension is used in two biological databases (wikis): CHDwiki and YTPdb.


Annotation: Blog post presenting the a gamification input for Wikidata: Wikidata — The Game.


Annotation: Description of content migration from Google Freebase to Wikidata with the Primary Sources Tool.


Annotation: Built and evaluate a system for prediction of external links within a Wikipedia article.

**Annotation:** Describes IntelliGenWiki a MediaWiki-based wiki with natural language processing (NLP). The data is represented via a Semantic MediaWiki extension. The NLP is handled with a General Architecture for Text Engineering (GATE) pipeline and may, e.g., identify organisms and enzymes.


**Annotation:** Describes a system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes the system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes IntelliGenWiki a MediaWiki-based wiki with natural language processing (NLP). The data is represented via a Semantic MediaWiki extension. The NLP is handled with a General Architecture for Text Engineering (GATE) pipeline and may, e.g., identify organisms and enzymes.


**Annotation:** Describes a system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes IntelliGenWiki a MediaWiki-based wiki with natural language processing (NLP). The data is represented via a Semantic MediaWiki extension. The NLP is handled with a General Architecture for Text Engineering (GATE) pipeline and may, e.g., identify organisms and enzymes.


**Annotation:** Describes a system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes the system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes a system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes the system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes the system that integrates natural language processing with a MediaWiki-based wiki. NLP service is provided by a W3C standard web services running GATE NLP software. The system read data from a wiki and embed the result in the wiki. The system is demonstrated on several data/wiki: DurmWiki with cultural heritage data where the system extract terms for an index, ReqWiki for quality assurance in software requirements specifications on a wiki and GeneWiki with extraction of biological terms.


**Annotation:** Describes the SemperWiki semantic personal wiki that is implemented in Java.


**Annotation:** describes the SHAWN Wiki, a semantic wiki implemented in Perl with semantic input in the wiki-markup and a query feature.


**Annotation:** Describes the IkeWiki semantic wiki engine. It is implemented as a Java web application and has support for OWL-RDFS reasoning.


**Annotation:** Description of the SweetWiki semantic wiki engine. It does not use a wiki markup language but rather XHTML and a WYSIWYG interface based on the Kupu editor. It relies on the CORESE semantic search engine.


**Annotation:** Suggest a semantic wiki based on the MediaWiki software.


Annotation: Describes an extension to Semantic MediaWiki so the revision history of a MediaWiki is available to the semantic query system.


Annotation: Describes a student project that built a question answering system, http://askplatyp.us, with the use of natural language processing and Wikidata data.


Annotation: Short description of the DynaTable MediaWiki extension for structured data in tables.


Annotation: Describes a wiki for supporting extreme programming. The wiki can associate forms with wiki pages.


Annotation: Describes a wiki with form-based input applied for annotation of scientific papers. The user may modify how the form is presented for display and for editing. XML and wiki markup combine in a language to define the form.


Annotation: Describes a student project that built a question answering system, http://askplatyp.us, with the use of natural language processing and Wikidata data.
Annotation: Describes the WikiCreole language that is a standardized wiki markup language.


Annotation: A blog post on the Wikimedia Foundation Parsoid project that creates a HTML5 representation of MediaWiki wikitext.


Annotation: News article on the controversy surrounding the VisualEditor on Wikipedia.


Annotation: Swedish introduction to Wikisource and its inspiration from Runeberg.


Annotation: Describes a system for collecting texts from the Arabic Wiki-source to form a corpus that can be used in a text mining application for plagiarism detection.


Annotation: Editorial on volunteered geographic information systems


Annotation: Introduces Open-StreetMap with discussion of among other issues the historic development, input and output methods, technical infrastructure, distributed map tile rendering, social collaboration, mapping parties, motivations, data accuracy and participation inequality.


Annotation: Patent on some aspects of histropedia.


Annotation: Description of ‘Copernicus’: A encyclopedia/browser that combines Wikipedia article reading with a 3-dimensional viewer in a two-layer interface paradigm. The article is presented as a transparent foreground, the 3D model in the background.


Annotation: Fairly briefly describes EtherPad with wiki and other extensions.


Annotation: Fairly briefly describes EtherPad with wiki and other extensions.


Annotation: News article about Ward Cunningham’s Smallest Federated Wiki.


Annotation: Guidelines for scientists on how to write on Wikipedia.


Annotation: Survey among university faculty for use of Wikipedia and attitude towards Wikipedia’s credibility, as well as a content analysis of the use of Wikipedia in research papers.


Annotation: Short comment that encourages scientists to contribute to Wikipedia.


Annotation: Short article encouraging ornithologists to contributed to Wikipedia. Professors could also ask students to create or improve Wikipedia taxonomic articles.


Annotation: Short article encouraging healthcare professionals to get involved in Wikipedia contribution.

Annotation: Short article on the collaboration between Wikipedia editors and the Cochrane Collaboration.


Annotation: Examines how college students use (i.e., read) Wikipedia.


Annotation: News article reporting a study by M. Namdari et al., ‘Is Wikipedia taking over textbooks in medical student education?’ that investigated which information sources medical students uses when taking an psychiatry exam. Question books, Up-to-Date and Wikipedia were used the most.


Annotation: Discusses how students could use Wikipedia.


Annotation: Details a project were High School students wrote science wiki articles.


Annotation: Historian in Japanese history from Middlebury College describes the students of his using erroneous information on Wikipedia, the Wikipedia policy subsequently adopted by his department and the media coverage. He also makes general remarks on authority, accountability and validity.


Annotation: Middlebury College history department has banned students from using Wikipedia as a source.


Annotation: Blog article reporting an analysis of citations from academic journals in the humanities domain area to Wikipedia.


Annotation: Historian in Japanese history from Middlebury College describes the students of his using erroneous information on Wikipedia, the Wikipedia policy subsequently adopted by his department and the media coverage. He also makes general remarks on authority, accountability and validity.


Annotation: Middlebury College history department has banned students from using Wikipedia as a source.


Annotation: Blog article reporting an analysis of citations from academic journals in the humanities domain area to Wikipedia.


Annotation: Article describing a new approach taken by scientific journal RNA Biology where author are required to also submit a Wikipedia article.


Annotation: Presents the concept of "Topic Pages" in the scientific journal PLoS Computational Biology. It is educational academic articles published in the journal that are meant to also be published on Wikipedia.


based on the OmegaWiki MediaWiki extension.


Annotation: A blog post with a critical examination of the then just launched WikiProteins biology wiki.


Annotation: Spanish description of the WikiPapers wiki for structured annotation of academic papers about wiki research.


Annotation: Blog post on German court cases involving Wikipedia information and right of personality.


Annotation: Describes reputation management among scientists


Annotation: Statistics from poll on online reputation management by researchers


Annotation: Short article arguing for Wikipedia as an opportunity for teaching students about collaborative writing in digital environments.


Annotation: Reports on the experience with given Wikipedia writing assignment in a history course.


Annotation: Reports on a Comparative Law course where students were to write and review Wikipedia articles.


Annotation: Short report on how participants in a graduate seminar wrote Wikipedia articles.


Annotation: A review of students assignment in Wikipedia contribution as they were listed on Wikipedia.


Annotation: A short description of the use of a wiki for course communication.

[480] Norm Friesen and Janet Hopkins. Wikiversity; or education meets the free culture movement: An ethnographic investigation. First Monday, 13(10), October 2008.

Annotation: Describes the experiences from giving a course through Wikiversity.


**Annotation:** Survey among 80 contributors to the Wikibooks Wikimedia project as well as email interview with 15 individuals selected among the 80. Among the several results were that contributors/respondents tended to be young males.


**Annotation:** Reports on the extension and update of a textbook with wiki technology where students participate. The researchers found that for some classes the student participation was associated with higher grades on average.


**Annotation:** Study on the productivity of scientists by counting number of publications, finding a heavy-tailed log-normal distribution. Explanations for the phenomenon are put forward.


**Annotation:** Small study on programmer productivity


**Annotation:** Report on Justin Knapp that was the first to reach one million edits on Wikipedia.


**Annotation:** Describes a machine learning-based system for predicting future number of edits of Wikipedia contributors


**Annotation:** Describes a machine learning-based system for predicting number of edits a contributor make on Wikipedia.

[490] Piotr Konieczny. We are drowning in promotional artspam. The Signpost, April 2015.

**Annotation:** Op-ed on the problem of promotional articles on Wikipedia.


**Annotation:** Interview with Jimmy Wales about Wikipedia.


**Annotation:** Example on a peer-reviewed work published on a Wikimedia Foundation wiki. The small study weighed 19 different teaspoons and found an average weight on 25 grams.


**Annotation:** Theoretical considerations and simulations


**Annotation:** Critique of Ioannidis papers, e.g., the critique notes that Ioannidis is wrong when he only uses 0.05 for his simulation for his claim and not the actual p-value observed in the studies.

Annotation: This is the indepth critique of Ioannidis paper summarized on an article in PLoS Medicine.


Annotation: Study on whether previous highly-cited clinical research papers are replicated.


Annotation: A study of what effect the censoring of Wikipedia in China had on the development of the Chinese Wikipedia.


Annotation: News article on the troubling economy of Encyclopaedia Britannica.


Annotation: Study on reading comprehension with 43 Year 6 pupils where half was reading on texts on iPads.


Annotation: Danish news article about the result of the 2012 Programme for International Student Assessment (PISA) in Denmark and the score of pupils in regard to whether they have access to tablet computers or not.


Annotation: Article discussing Wikidata and its possible influence on the diversity of knowledge represented in Wikipedia.


**Annotation:** An investigation of 7058 contributors on the French and Dutch Wikipedias, to answer whether contributions from registered users have higher quality than anonymous users. Each user is measured wrt. retention of their added text, while not considering, e.g., edit wars. They find that the retention is higher for anonymous users.


**Annotation:** A book dubbing ‘Homo generousus’ and discussing issues such as altruism, gift economy, generosity, reciprocity, cooperation, moral, costly signals, the handicap principle, sexual selection and economic games.


[531] Philip Ball. The more, the wikier. news@nature.com, February 2007.


A Notes

There are numerous other studies on wikis, Wikipedia and related areas not summarized here. The following are some notes and discovered links not yet incorporated into the body text:

A.1 Publications

15. Jon Chamberlain, Udo Kruschwitz and Massimo Poesio. Constructing an Anaphorically Annotated Corpus with Non-Experts: Assessing the Quality of Collaborative Annotations
18. Empirical Analysis of User Participation in Online Communities: the Case of Wikipedia. Giovanni Luca Ciampaglia, Alberto Vancheri
26. Donghui Feng, Sveva Besana and Remi Zajac. Acquiring High Quality Non-Expert Knowledge from On-Demand Workforce
32. Zeno Gantner and Lars Schmidt-Thieme. Automatic Content-Based Categorization of Wikipedia Articles


39. Haigh. Wikipedia as an evidence source for nursing and healthcare students

40. Laura Hale (2013). User:LauraHale/Blocks on English Wikinews


45. A. J. Head and M. B. Eisenberg, 2010, How today’s college students use Wikipedia for course-related research

46. Heilmann et al.: Wikipedia : a key tool for global public health promotion


53. Jonathan Isbell, Mark H. Butler (2007). Extracting and Re-using Structured Data from Wikis


55. Richard Jensen, “Military history on the electronic frontier: Wikipedia fights the was of 1812”. Link


60. Maximilian Klein, Piotr Konieczny. Gender gap through time and space: a journey through Wikipedia biographies and the “WIGI” index.


70. Yinghao Li et al. (2007). Improving weak ad-hoc queries using Wikipedia as external corpus
72. Meyer et al. (2007). Categorizing Learning Objects Based On Wikipedia as Substitute Corpus. To help categorize unlabeled learning objects
73. Mietchen et al. 2011, Wikis in scholarly publishing
77. Milne, Witten. Learning to link with Wikipedia
78. Minol K, Spelsberg G, Schulte E, Morris N. Portals, blogs and co.: the role of the Internet as a medium of science communication.
82. Timme Bisgaard Munk (2009), Why Wikipedia: self-efficacy and self-esteem in a knowledge-political battle for an egalitarian epistemology
84. Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry and Chu-Ren Huang. Wiktionary for Natural Language Processing: Methodology and Limitations
86. Mathieu O’Neil, Cyberchiefs: Autonomy and Authority in Online Tribes.
89. Ayelet Oz, Legitimacy and efficacy: The blackout of Wikipedia.
90. Manuel Palomo-Duarte, Juan Manuel Dodero, Inmaculada Medina-Bulo, Emilio J. Rodríguez-Posada, Iván Ruiz-Rube (2012). Assessment of collaborative learning experiences by graphical analysis of wiki contributions
100. Colleen A. Reilly (2011), Teaching Wikipedia as a mirrored technology
101. Reinoso et al., A quantitative approach to the use of the Wikipedia.54
104. Scheider et al (2010). A qualitative and quantitative analysis of how Wikipedia talk pages are used (same as “A content analysis: how Wikipedia talk pages are used”)


121. Benjamin Mako Hill (2013). Essays on volunteer mobilization in peer production

122. Timothy Weale, Chris Brew and Eric Fosler-Lussier. Using the Wiktionary Graph Structure for Synonym Detection


130. Mark Yatskar. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia.


A.2 Further notes

- Aaron Halfaker (2013). Research:Wikipedia article creation
• Freebase, http://www.freebase.com/
  http://www.wikicfp.com/
  http://dbpedia.org/sparql
• Ontology construction SUMO, OpenCyc, Yago, UMBEL
• Google Squared and Google Fusion Tables.
• International Music Score Library Project
• http://debategraph.org/
• Searching Wikipedia WikiWax “Semantic Wiki Search”.
• Wikipedia as a teaching tool workshop.odp
• WikiXMLDB http://wikixmldb.dyndns.org is an XML database using Sedna
• Visualization researchers have used the many kinds of statistics extracted from Wikipedia in large-scale renderings of network indegrees. http://scimaps.org/maps/wikipedia/20080103/
• The collaborative debugging tool http://pastebin.com/ with syntax highlighting.
• http://www.mkbergman.com/?p=417
• WikiMedia Foundation has maintained a computer system with fast response time. BSCW was slow
• User authentication
  “real name” legislation in South Korea. Link
• Research to the people
• Web service for generation of SIOC files for Medi- aWiki page
• Survey of tools for collaborative knowledge con- struction and sharing.
• ... and wikis and Wikipedia have been the subject of many theses.539
• Game theory: contributing to the common good increase the value more than the contribution. Other apparent irrational behavior: voting. Performance on a work sample test is among the best predictor for job performance.540
  “people like to see their words in print”196
• Xanadu
• Poor Man’s Checkuser
• Visualizations: Link
• ZooKeys, Species ID wiki
• Encyclopedia of Earth, Environmental Health Per- spectives
• Useful Chemistry
• Redaktionen editorial team
B  Further links

http://figshare.com/articles/Wikipedia_Scholarly_Article_Citations/1299540
http://libresoft.es/Members/jfelipe/thesis-ukp-quantanalysis
NOAM COHEN, Don’t Like Palin’s Wikipedia Story? Change It,
Wikilink manipulation.
http://www.bluehost.com/how-to-overthrow-a-wikipedia-result/
http://grey.colorado.edu/knight/index.php/Wikipedia_Quality_Pagerank_Correlation_Coefficient
http://www.economist.com/search/PrintFriendly.cfm?story_id=1144606
Amichai (Hamburger, Landez, Madiel and Hayat in 2008
Kevin Clason receiving hate mail.
Vandalism?
http://www.zipom.com/versions/wikipedia/home.html
'In case anyone is interested, we tried to statistically map freebase types to dbpedia types, the key ideas are described here:
A variant of this paper was just accepted to the International Semantic Web Conference (ISWC).'
Kevin Clason receiving hate mail.
http://wikimania2007.wikimedia.org/wiki/Proceedings/4D1
http://meta.wikimedia.org/wiki/Editor_by_project_and_country_of_origin
http://mako.cc/copyrighteous/20100627-00
http://meta.wikimedia.org/wiki/Mind_the_Gap
Suggestbot User:SuggestBot
Suggestbot: using intelligent task routing to help people find work in wikipedia
wikiFeed http://wikistudy.mathcs.carleton.edu
Index

Æblerød, 17

addiction, 30
altmetrics, 36, 37
altruism, 29
American Customer Satisfaction Index, 26
article trajectory, 17
associativity, 17
augmented reality, 37
authorship, 21

barnstar, 18, 29, 30, 53
bazaar, 24
benevolent dictator, 24
Bicholim conflict, 17
biclique, 16
bot, 15, 39, 40
Brede Wiki, 43, 44, 49
Broadwell, Paula, 35
certification, 39
CheiRank, 37
citation, 3, 47, 48
ClueBot, 19, 38, 39
ClueBot NG, 19, 38
clusterball, 16
Cochrane Collaboration, 47
collaborative writing application, 6, 7
comScore, 25
constitutional monarch, 24
court case, 14, 15
coverage, 10–13
CoWeb, 50
critical mass, 24
cross-cultural study, 16
CrossRef, 13

DBpedia, 1, 37
distributed wiki, 45
donation, 28
economy, 27, 28
edit-a-thon, 27
egoboo, 17, 30
election, 36
Encarta, 7, 8, 32, 55
Encyclopædia Britannica, 7, 8, 11–14, 48, 55
Everything2, 14

Facebook, 3, 20, 26
findvej.dk, 37
Finkelstein, Seth, 24, 29, 53
Firefox, 44
flagged revision, 32
folksonomy, 34
formality, 14
free and open source software, 23, 24
free culture, 6

Freebase, 41
gamification, 41
Gardner, Sue, 15
gender, 10, 12, 16, 20, 22, 23
Gene Wiki, 25, 40
General Inquirer, 20
genre, 14
geotagging, 24
GermaNet, 9
Gini coefficient, 21
Gompertz model, 31
Google, 34
Google Knol, 32
Google Scholar, 2, 25
Google Trends, 25
Googlepedia, 44

health, 36
Heilman, James, 49
HistoryViz, 44
histropedia, 44
hoax, 9, 17
HTML5, 43
hypertext, 32

image retrieval, 33
index generation, 41
infobox, 16, 23, 35, 40, 43
IntelliGenWiki, 41
Internet penetration, 52
IRC, 4

Javascript, 34
Kaggle, 53
Kane, Gerald C., 39

labor hours, 27
LaTeX, 52
Layar, 37
Legolas2186, 17
license, 21
Linked Data, 1
Linked Open Data, 42
Linux' Law, 32
Linux, 4, 12, 28
List to Wikipedia, 35
LOINC, 10

machine learning, 5, 19, 33, 38, 39
Macropedia, 13
Manske, Magnus, 12, 40–42, 56
markup, 43
MD5, 39
MediaWiki, 37, 40
Metaphacts, 40
Meyer, Bertrand, 39
Microsoft, 19
Mix’n’match, 12
Monkey selfie, 15
motivation, 28–30

named entity recognition, 33
natural language processing, 41
network, 17
network analysis, 15–17
network visualization, 16

oh shit graph, 31
OmegaWiki, 49
One Laptop Per Child, 45, 55
online community, 6
Open Access Reader, 54
OpenThesaurus, 9
OpenWetWare, 49
opportunity cost, 28

page view, 36
page views, 3
PageRank, 15, 16, 37
Palin, Sarah, 35
Parsoid, 43
PDBWiki, 49, 54
persistent word view, 21, 22
personality right, 49
PISA, 55
policy, 53
popularity, 25, 27
power law, 15
prediction, 35, 36
Prefuse, 16
Primary Sources Tool, 41
productivity, 53
PyPedia, 46
Python, 46

qLabel, 34, 56
quality, 6, 7
query expansion, 33
question-answering, 42

Ramsey, Derek, 39
Raymond, Eric, 24, 30, 32
readability, 7–9
Reasonator, 40, 41, 44
recommender system, 16
reification, 42
replacement cost, 27
reputation, 29
review, 2, 33
revision history, 38
Rfam database, 18, 40, 49
RNA Biology, 48

Scholarpedia, 32, 49
Semantic MediaWiki, 42, 43, 49
semantic relatedness, 33
semantic stability, 40
Semantic Web, 42

semantic wiki, 5, 41, 42
sentiment analysis, 17, 20
Slashdot, 26, 32
SNPedia, 49
social movement, 6
sockpuppet, 20, 53
Spanish Fork, 27, 31
SPARQL, 42
Special:Nearby, 41
stigmergy, 25
STiki, 19, 38, 39
Summers, Ed, 4, 13
Swartz, Aaron, 21, 32

ToolsServer, 3–5, 18, 23, 35
transaction cost, 32
translation, 34
trust, 38
Twitter, 6, 13, 36

up-to-dateness, 7, 8, 13
Usenet, 31

vandalism, 9, 17–19, 37–40, 51, 56
Veropedia, 39
VisualEditor, 28, 43

Wales, Jimmy, 12, 21, 24, 47, 54
West, Andrew, 35
WhoKnows, 37
Wiki ShootMe, 41
Wiki Trip, 24
Wiki-PR, 28
WikiBlame, 21
Wikibooks, 47, 52
WikiBrowser, 45
WikiBu, 39
WikiChanges, 35
WikiChecker, 23, 35
WikiCite, 3
WikiCreole, 43
WikiDashboard, 21
Wikidata, 3, 6, 12, 13, 34, 35, 37, 38, 40–43, 55, 56
Wikidata Query editor, 42
Wikidata Query Service, 42
Wikidata Toolkit, 4
Wikiganda, 20
WikiGenes, 49
wikiGraph, 45
wikilink, 15, 16

Wikimedia Commons, 34
Wikimedia Tool Labs, 3, 4, 35
WikiOpener, 40
WikiPathways, 49
Wikipedia, 19
Wikipedia Beautifier, 44
Wikipedia Live Monitor, 35
Wikipedia Natural Disaster Monitor, 35
Wikipedia Recent Changes Map, 24
WikipediaVision, 24
WikiPic, 34
WikiProtein, 54
WikiProteins, 49
WikiTage, 35
WikiScanner, 19, 36
WikiSource, 43, 44, 47
WikiStory, 44
WikiStream, 4
WikiTravel, 31
WikiTrends, 35
WikiTrivia, 37
WikiTrust, 20, 38, 39
WikiTude, 37
wikitweets, 13
WikiUniversity, 51, 52, 54
WikiViz, 16
WikiWord, 34
WiKtionary, 9, 33
WiViVi, 24
WordNet, 33, 35
WYSIWYG, 43

XOWA, 45
XWiki, 45

YAGO, 33, 35

Zachte, Erik, 5, 24, 31