Researching Wikipedia - Current approaches and new directions

Phoebe Ayers
University of California, Davis Physical Sciences and Engineering Library
One Shields Avenue University of California Davis, California 95616-8676
psayers@ucdavis.edu

Wikipedia (<http://www.wikipedia.org>), an international, multi-lingual and collaboratively produced free online encyclopedia, has experienced massive growth since its inception in 2001. The site has become the world’s single largest encyclopedia as well as one of the world’s most diverse online communities. Because of these factors, the site provides a unique view into the processes of collaborative work and the factors that go into producing encyclopedic content. To date, there has been no unified review of the current research that is taking place on and about Wikipedia, and indeed there have been few formal studies of the site, despite its growing importance. This project is a review of social science and information science studies of the site, focusing on research methods and categorizing the areas of the site that have been studied so far. Studies of Wikipedia have focused primarily on the social dynamics of contributors (such as how disputes are resolved and why contributors participate), and the content of Wikipedia (such as whether it is an accurate source), but due to the unique collaborative processes on Wikipedia these two areas are deeply intertwined.

Introduction

Wikipedia (<http://www.wikipedia.org>) is an online, free-content international and multilingual encyclopedia, operated by the Wikimedia Foundation (http://wikimediafoundation.org). The site is collaboratively produced by volunteers, and is based on wiki software, which allows contributions from anyone with a web browser and internet access. Contributors may be anonymous, and nearly every page of the site is editable. Wikipedia has experienced massive growth since its inception in 2001, becoming one of the world’s largest reference works, with over a million articles in the English language Wikipedia alone. The contributors represent one of the world’s most
diverse online communities, with projects established in over two hundred languages. Because of these factors, the site provides a unique view into the processes of collaborative work, demonstrating how people brought together with a common goal can produce meaningful results, develop community guidelines and policy, and resolve differences. Wikipedia is also unique as a project in producing “encyclopedic” content with no top-down editorial control or mandatory peer review processes. This raises questions and concerns about the accuracy, validity and scope of the site’s content. These questions seem to have not impacted the high use of Wikipedia, however, which as of May 2006 was ranked as one of the top twenty websites in the world by use.

Wikipedia thus offers multiple opportunities for interesting research projects. To date, there has been no review of the research that is taking place on and about Wikipedia, and indeed there have been few scholarly studies of the site, despite its growing importance and inherent interest. While there has been much commentary about Wikipedia in the popular and scholarly press alike, there has been little attempt to systematically examine the site. Several master’s theses and PhD dissertations are underway, but there is little published work. In part, this may be due to the newness of the site, and in part to the difficulties of studying Wikipedia — studying anonymous, open, rapidly changing wikis present unique challenges to the Internet researcher.

This paper is a review of social science and information science research studies done on and about Wikipedia. While there’s a larger literature of studies done on wikis in general, this review focuses exclusively on scholarly studies that are primarily about or based on Wikipedia. This paper also outlines research questions that have been raised to date and suggests future directions. It is hoped that examining research endeavors undertaken so far and particularly focusing on study methods and methodologies will help inform future research on the users, communities, content and technical infrastructure of Wikipedia.

This review is necessarily incomplete; there are a number of studies that mention Wikipedia as a case study that are not included which may be valuable for understanding the site. There is also a rich literature in German and other languages that has not been fully addressed. This study also does not address the technical aspects of the site (such as the scalability of the site or the possibility of adding a semantic data structure to MediaWiki), some of which are being researched and many of which are worthy of further study (see Völkel et al. 2006). Furthermore, research on wikis in general, and on Wikipedia and Wikimedia Foundation projects specifically, is just now beginning to blossom. There are numerous projects that are currently in development, as can be seen by looking at the conference programs for WikiSym 2006.
(an ACM-supported conference on wikis) and Wikimania 2006$^9$ (the annual conference of the Wikimedia Foundation) both of which are occurring in August 2006 and are not represented in this paper.

Research questions

There are two broad types of questions about Wikipedia that sociological and information science studies of the site have addressed. First, researchers may seek an understanding of the content of Wikipedia - for instance, whether it is reliable and accurate, how articles develop, and what the scope of the encyclopedia is; and second, they may seek an understanding of the community of Wikipedia - for instance, how consensus is reached in disputes, and what motivates participants. Because of Wikipedia’s unique production processes, both of these types of questions are deeply interrelated, and may be addressed in a single study. For instance, how contributors interact on an article may affect the content of that article, and the type of article (whether it is a controversial topic, or a policy page, for instance) may affect how collaboration occurs and who participates in editing. Within these broad areas of content and community there are many specific research questions, as indicated above.

Methodologies

Examining the studies of Wikipedia done so far in terms of research methods gives insight into how researchers can approach studying the phenomena associated with collaborative and anonymous communities, and the work that remains to develop a theoretical and methodological framework for studying these types of dynamic applications. While there has been written on methods for doing Internet research (see for example, Best and Krueger 2004, Hine 2005), studying wikis pose special problems. It is difficult, for instance, to gather a valid sample of contributors or users of open wikis with a method that remains unobtrusive and respectful of privacy. When looking at Wikipedia in particular, the sheer amount of data involved in looking at page histories and the rapidity of change of the site is daunting for content analyses. Furthermore, understanding the labyrinthine Wikipedia community and how it functions can be difficult for an outsider.

However, the technical features of MediaWiki (the software that runs Wikipedia), and the social and community structure of Wikipedia also provide opportunities for the researcher. In particular, the availability of complete edit histories for articles, logs of administrative actions such as deletions and protections, database dumps of changes
made to the site, discussion pages and user pages, and extensive archived mailing lists all provide rich data sources for research. Understanding the information that each of these sources of data is able to provide, and how they combine to provide a complete picture of how Wikipedia works, is a major challenge for the prospective Wikipedia researcher. Additionally, recognizing that no one source of data is likely to give a truly accurate picture of community or content- for instance, the people who heavily participate on mailing lists may not be the same people who participate in other areas of the site - is key. In the review that follows, studies are organized first by whether they primarily focused on Wikipedia content or community, and then by what aspect of the site they primarily studied (e.g., talk pages, article histories, mailing lists, etc.). Table 1 gives an overview of these areas of the site, a brief description, and whether the study done was primarily qualitative or primarily quantitative.

Table 1. Areas of Wikipedia addressed by reviewed studies

<table>
<thead>
<tr>
<th>Area of Wikipedia</th>
<th>Description</th>
<th>Quantitative studies</th>
<th>Qualitative studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article edit histories</td>
<td>The record of every change made to a page, with authorship (username or IP address), date and the author’s edit summary</td>
<td>Lih (2004), Viégas et al (2004)</td>
<td></td>
</tr>
<tr>
<td>Site statistics</td>
<td>E.G., statistics on growth of various Wikipedias over time, available for every language</td>
<td>Voss (2005)</td>
<td></td>
</tr>
<tr>
<td>User “talk” pages</td>
<td>Like article talk pages, attached to a user page and intended for discussion between users.</td>
<td>Voss (2005)</td>
<td></td>
</tr>
<tr>
<td>Policies and policy pages</td>
<td>Special pages that describe Wikipedia policies and guidelines. Most are open to editing, like articles.</td>
<td>Reagle (2004), Matei and Dobrescu (2006)</td>
<td></td>
</tr>
<tr>
<td>Article content</td>
<td>The actual content of Wikipedia articles, supposed to be NPOV, verifiable and not under copyright.</td>
<td>Brändle (2005), Emigh and Herring (2004), Holloway et</td>
<td>Giles (2005), Emigh and Herring (2004),</td>
</tr>
</tbody>
</table>
Researching Wikipedia content

Examining the dynamics of article creation and growth, particularly through article histories, has proved a particularly fruitful way to study site content and its change over time. Article histories are a record of every change that has been made to the site, and include authorship information (either a registered username or an “anonymous” IP address); date and time of the change, a link to the saved version that includes the change, a link to the differences between the saved version and the one previous to it, as well as to the current version; and an author’s (optional) summary of their changes. Examining the article history is also one possible way to examine article quality. Lih, in a 2004 study, developed quantitative metrics for developing a “reputation” benchmark of articles based on the number of edits (rigor) and number of contributors to each article (diversity), assuming that more rigor and diversity means higher quality articles. He used this benchmark to analyze the reputation of Wikipedia articles that were cited by the press, before and after they were cited. He found that press citation often led to increased editing (and thus presumably higher quality). Viégas et al (2004) study the edit history of articles with a visualization tool called History Flow, which measures how page content varies over time by individual edit. This visualization can reveal patterns such as edit reversions, edit wars, and mass deletions and additions. They use the visualizations produced by this tool to analyze how content in articles is produced and whether it stabilizes, and to determine how the community functions. They find that vandalism (which they find five types of) to articles is quickly removed, that there is large
variation between articles in the number of anonymous edits they receive, and that there is no clear link between anonymous editors and vandalism. They also find that page size does not stabilize over time for most articles, a pattern that has implications for both the content accuracy and community dynamics of Wikipedia.

Whether the actual content of articles is accurate and verifiable is also a question of much concern. Content may also be examined in terms of writing style and formality, and whether it is factually accurate and complete. Giles (2005), in a comparative analysis of Wikipedia and Encyclopaedia Britannica published in Nature magazine, used expert reviewers to review a selection of fifty articles on science topics, and compared the number of errors that were found in each to produce a comparison of the two encyclopedias. He found that while Wikipedia articles suffered from poor writing, their rate of error was comparable to Britannica’s. Emigh and Herring (2004) use quantitative discourse analysis and qualitative observation to determine how articles from collaborative online systems such as Wikipedia and Everything differ from traditional print encyclopedias, focusing on the degree of formality in language use. Their discussion focuses on how the different types of editing afforded by the sites result in different types of articles, finding that Wikipedia articles - which are produced in a system that allows a great deal of what they term “post-production editorial control” - have a formal tone, coming close to the tone of traditional encyclopedias such as the Columbia Encyclopedia. They conclude that open editing systems can result in stylistic homogeneity.

Brändle (2005) uses a quantitative content analysis to research the question of what factors generally lead to quality in Wikipedia articles, using the perspective of communication and media studies. Brändle proposes a "topic-attention-quality" model, which has quality as a dependent variable; quality is defined in terms of a number of facets including references, layout, whether the article answers the questions of who, what, where and when, and so on. He reviews 450 articles from the German Wikipedia in terms of this model, studying the effects various measures have on quality. A factor analysis revealed five factors: relevance, interest, controversy, richness, and neutrality. Brändle found that relevance (the newsworthiness of a topic, its representation in Brockhaus (a German encyclopedia) and Google hits), and interest (a measure of the traffic to and edit history of an article) together account for 64% of the quality variable of richness. Stivila et al (2005a) do a similar study on information quality of Wikipedia articles, in which they quantitatively analyze the article histories for 834 randomly chosen articles and the set of 236 Featured Articles that existed at the time of their study. They develope 19 descriptive metrics that may have a role in the information quality of Wikipedia articles, including such measures as total number of edits, number of anonymous user edits, and age of the article. They then used factor analysis to get
seven information quality metrics: authority/reputation, completeness, complexity, informativeness, consistency, currency and volatility. They analyzed article content in terms of these factors, using Featured Article measurements as benchmarks for quality. Through statistical analysis, they find that these benchmarks are useful for discriminating high-quality articles, thus possibly providing an automatic way to detect high-quality content.

Content may also be examined in terms of scope - whether Wikipedia adequately covers a given topic or domain. There has long been concern over Wikipedia’s scope in traditional ‘scholarly’ topics, though it is generally recognized as having very good coverage of other types of topics, such as popular culture or computer topics. Wikipedia’s scope coverage may be examined in terms of a domain or subdomain. For instance, Altmann (2005) surveys Wikipedia’s coverage of the topic of medical informatics, finding that the site is lacking several important topics in this area.

Wikipedia content is comprised of more than simply article text, however; for instance, categories may be added to an article, either by hand or by the addition of a template (such as a stub message). Holloway et al. (2005) study the semantic coverage of topics in Wikipedia by mapping the organization of categories and subcategories across articles in the English-language Wikipedia, a year and a half after they were first introduced (in May 2004). They found that as of November 2005, roughly half of all articles had no categories. (They also analyze other statistics, finding for instance that nearly half of the unique authors in the November 2005 dump were registered but had made only one edit). They then developed a a measure of similarity of categories, defining similarity as whether two categories were used in the same article or not, and map similar categories using a visualization tool, showing that similar categories occur as logical clusters. They also map the last edit time of categories and the top ten authors, finding that categories are mostly kept current.

The discussion pages of articles also provide a framework for understanding both collaborative processes and how the content of an article develops. Stvilia et al. (2005b) qualitatively analyze the information quality (IQ) of Wikipedia articles and the processes and discussions that by which quality is developed in articles. They use a 21-point framework to determine IQ, comparing this to other measures such as the Featured Articles guidelines in Wikipedia. They then use this framework to analyze the points brought up in the discussion pages of a sample of 834 randomly selected articles and 236 Featured Articles (all that existed at the time of their study). They use this analysis to determine what measures of IQ contributors use to determine quality, and what problems exist on these articles, such as a lack of verifiability or readability. They also use this analysis to study how participants cooperate and compromise on content, and
what trade-offs are required when writing an article (between completeness vs accessibility, for instance). They also analyze Featured Article Removal Candidates (FARC)s for the elements of quality (or lack thereof) that are discussed during the removal process.

Statistical analysis of the site itself - the growth of Wikipedia across languages, for instance, and the growth of articles - is also possible, by analyzing the log files and data dumps of changes made to the site that are available. Voss (2005) discusses possible research that could be done with these statistics. For example, he compares the growth rate of Wikipedias, (finding that there is a linear and an exponential growth phase for all Wikipedias); the size of namespaces and ratio of user talk pages to user pages and article talk pages to articles in Wikipedias across a handful of languages (finding that these ratios vary across languages); compares article size distribution (finding that they are lognormally distributed); and link structure; and graphs contributors per article and edits per contributor (finding that in the German Wikipedia nearly half of all articles have less than five distinct authors, and a third of all authors have only contributed to one article)

Researching the Wikipedia community

Two studies examine conflict resolution and decision making processes on Wikipedia by examining mailing list communications and the discussion pages of articles. Reagle (2004) examines email lists, talk pages, and policies, including Neutral Point of View (NPOV) in light of a review of the “interdependent decision-making literature” in order to examine Wikipedia’s consensus processes, finding that it is complex and and sometimes difficult to distinguish between agreement and disagreement. Matei and Dobrescu (2006) qualitatively analyze one Wikipedia mailing list thread and the discussion page of one key policy on Wikipedia (NPOV) to analyze conflict and policy discussion among contributors, arguing that the many conflicts and debates on Wikipedia and its communication channels such as mailing lists lead to meaning pluralization and fundamental ambiguity of key policies such as NPOV.

One technique for finding out the motivations of contributors and what they think of the site is to directly ask them. Bryant, Forte and Bruckman (2005) use participant interviews to qualitatively characterize the changing participation of contributors to the site from a “novice” level to an “expert” level. They interview nine established and active contributors, recruited on Wikipedia, to determine how contributors began participating in the site and how their participation and how they see themselves as contributors changed over time. The paper uses two social theories, activity theory and legitimate
peripheral participation, to contextualize their findings that contributors begin by editing limited subjects and pages, and slowly turn into contributors who take responsibility for different areas of the site's operation beyond simply editing articles.

 Broadly surveying contributors in order to find out their motivations is also a technique that has been proposed. One such survey was done by Schroer and Hertel (2005), who have not yet published their results. A similar study has been proposed on the Meta Wikimedia site but not yet implemented. There have been several informal onwiki surveys to determine contributor demographics, such as an effort to plot the ages of contributors on the German Wikipedia, but these are typically undertaken without efforts to produce a statistically valid sample.

 Additionally, users on Wikipedia may choose to provide publicly available information about themselves on their “user page.” These pages may state personal demographics, interests, or beliefs, as relevant to participant's activities on or off the wiki. While user pages are widely used, and their content has been the subject of much debate on Wikipedia, they have been relatively understudied. Cathy Ma (2005) qualitatively analyzed 100 random user pages on a 24-construct scale in order to determine what information is disclosed by those who contribute to Wikipedia, and determine whether disclosure was related to the number of edits a contributor made, finding that there are key disclosure variables that do link to a contributor's overall participation. Voss (2005) compared the ratio of user pages to user talk pages across different language Wikipedias, finding that this ratio varied by language (the Japanese, for instance, seem to have more user talk pages than user pages).

 Another avenue is quantitatively studying contributor data, which is available in the form of edit logs for each registered username or unregistered IP address - that is, every change that a username or IP makes is available in an edit history. Voss (2005) studies the distribution of distinct articles edited per author in the German Wikipedia, finding that a third has only ever contributed to one article. Similarly, a recent unpublished study sampled edits from the English Wikipedia since 2002, and then looked at the date of account creation for each of the contributors associated with these edits, tentatively finding that more than 25% of all edits were done by accounts that had been registered less than 100 days, an indication that new users play a large role in the site's content. Stvilia et al (2005a) determined that for their pool of 834 randomly chosen articles, 6% of the editor pool did 24% of the edits, while for Featured Articles the ratio is even lower - 2% did 21% of the edits.

 There is also an ongoing debate about the value of “anonymous” contributions versus
registered user contributions. Anecdotally, anonymous users are often responsible for vandalism on the site. Anthony et al (2005), in an unpublished study, use statistical analysis of the percentage of an individual contributor’s edit retained between versions of an article as a measure of “quality” of a particular contributor’s contributions. Using this difference measure, they measured the quality of 7083 anonymous and registered contributors to the French and Dutch language Wikipedias, and then compared anonymous to registered accounts. They found that while registered contributors have a higher level of quality edits than anonymous contributors at a high level of total edits (>5), surprisingly, anonymous contributors have a higher level of quality edits than registered users at a low number of total edits (<4).

**Future research**

Research on wikis and on Wikipedia in particular is still in its infancy. While there are a number of proposed studies in process and underway, there are still a number of possible research questions that remain unaddressed. Further research questions about site content that have broad implications for the site and its mission include whether articles steadily improve over time; whether the content of articles is verifiable, and what kinds of sources (by type, or accessibility) are used to document articles, as well as what the scope and coverage of Wikipedia articles is. Questions involving the Wikipedia community include what the demographics and expertise of Wikipedia contributors are like; what the long-term participation history and contribution patterns of contributors are; and how contributors work together on specific projects on the wiki. Cross-project questions include how different language Wikipedias compare in growth and policies.

There are also unaddressed questions regarding outside use of the site and the current role and use of Wikipedia as a reference work. Questions that remain formally unstudied include examining users’ perceptions of and information behavior on Wikipedia. As suggested by Ayers (2005), such questions include why users (non-contributors) go to Wikipedia, what do they expect to find there, and how they use the site. How does this behavior vary between groups of users, such as reporters or librarians? Other work could examine citation patterns of Wikipedia to examine acceptance in various domains; Lih (2004) begins such an analysis of news stories but there are other domains that could be studied. Analyzing how quickly a Wikipedia in a particular language develops authority, or becomes a cited reference resource in that language, could be particularly interesting. Finally, usage patterns of Wikipedia - how people use the site internally and navigate from link to link - remains unstudied. Other research questions and proposed studies are being collected on the English Wikipedia Wikidemia project.
There are also a variety of other possible data sources from Wikipedia that could be used and analyzed besides the ones mentioned here, including community and “help” or question-answering pages; project pages (collective efforts to work on a particular task or topic area within Wikipedia are often broken out into projects); the links and linking patterns between articles; and administrative logs (such as protection or deletion logs). Additionally, the data sources that are analyzed in the studies reviewed here, such as edit histories, site growth logs, and article content, are extremely rich and would benefit from further study.

**Summary and conclusions**

Wikipedia, the free online encyclopedia, offers extraordinary opportunities for research into collaborative work processes, the motivations of those participating in open-source and free-content projects, and the quality of user-generated encyclopedic content. Despite enormous and rapid growth since the site's inception in 2001, and the growing importance of Wikipedia as a reference source and internet phenomenon, there have been few published scholarly studies of the site. Research questions that have been addressed focus on two distinct areas, the content of the site and the Wikipedia community. However, due to Wikipedia’s collaborative work processes, these areas are deeply entwined and affect each other in ways that may be unique to the site. This paper reviews and analyzes 18 recent studies focusing on Wikipedia in terms of their methods, the data source that they used, and their findings.

Several studies that study Wikipedia content attempt to develop metrics for measuring the quality of content and the factors that produce high-quality content. In particular, more editing of an article, perhaps driven by more traffic to that article (Lih 2004; Brändle 2005) seems to result in higher-quality articles. However, while it is possible to determine the quality of Wikipedia articles (Emigh and Herring 2004, Brändle 2005, Giles 2005, Stvilia et al 2005b), the other factors that drive quality are less easily explicated. Studies of contributors to Wikipedia affirm that, while as with other open-source software projects, a few people contribute to many articles while many people contribute to just a few (Voss 2005, Stvilia et al. 2005a), drawing conclusions about the value of anonymous and ‘newbie’ contributions may be misleading (Anthony 2005).

These studies cover a great deal of ground in terms of what area of the site they chose to investigate and the methods that they use; however, there is much other research that could be done. Only a few of these studies explicitly acknowledge the complex interactions that occur between Wikipedia community and content as articles develop.
Despite having a cohesive mission (to create free encyclopedic content) and having a number of shared community values that are often expressed in policies (such as NPOV), the individuals that contribute to Wikipedia represent an extremely diverse (linguistically and culturally) community, which is rarely acknowledged in these studies. Voss (2005) examines the possible differences between different language Wikipedias, and other studies look at the different roles that contributors may play (Viégas 2004, Bryant et al 2005, Stvilia et al. 2005b); but a deep analysis of the various ways that individuals participate in Wikipedia has not, to my knowledge, been done.

The Wikipedia project offers numerous possible avenues of research and a wide variety of data sources which are regularly captured and available for analysis. Research on wikis and Wikipedia is just beginning, and there are numerous areas that remain unstudied or that could be studied in further depth. It is hoped that this article will help stimulate ideas for further research involving data sources available from Wikipedia.

Notes

2 See http://meta.wikimedia.org/wiki/Wikipedia_is_more_popular_than ... for the latest Alexa rankings and an explanation of how sites are ranked.  Back
3 A bibliography of research done to date is being collected here: http://bibliography.wikimedia.de/ ; there is also a literature review section at http://meta.wikimedia.org/wiki/Wiki_Research_Bibliography  Back
4 http://meta.wikimedia.org/wiki/Research has a list of people pursuing theses and dissertations about Wikipedia. Much of which is collected in the Wiki Research Bibliography.  Back
6 http://www.wikisym.org  Back
7 http://wikimania2006.wikimedia.org  Back
8 The software that produces statistics from the database dumps and log files was written by Erik Zachte, and some of these statistics can be found at the Wikipedia statistics page -- http://stats.wikimedia.org/ . Database dump downloads can be found at http://download.wikipedia.org .  Back
9 There is also an ongoing quantitative research project that is primarily occurring on the English Wikipedia (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikidemia/Quant ), which is still formulating research questions and collecting data.  Back
10 http://meta.wikimedia.org/wiki/General_User_Survey/Questionnaire  Back
Back

Original graph creation by Tim Starling; initial analysis by Jakob Voss; based on questions posed by Jimmy Wales. See http://wm.sieheauch.de/?p=44 for more information. Back

See http://wm.sieheauch.de/?p=44 for more information. Back


References


Germany  http://meta.wikimedia.org/wiki/Transwiki:Wikimania05/Paper-CM1


A Case of Mutual Aid: Wikipedia, Politeness and Perspective Taking  

http://wy2x05.psychologie.uni-wuerzburg.de/ao/research/wikipedia.php?lang=en


http://mailer.fsu.edu/~bstvilia/papers/qualWiki.pdf


http://eprints.rclis.org/archive/00003610/