Creating, Destroying, and Restoring Value in Wikipedia

Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam,
Katherine Panciera, Loren Terveen, John Riedl
GroupLens Research
Department of Computer Science and Engineering
University of Minnesota
200 Union St. S.E.
Minneapolis, Minnesota 55455
{reid,jilin,lam,katpa,terveen,riedl}@cs.umn.edu

ABSTRACT

Wikipedia's brilliance and curse is that any user can edit any of the encyclopedia entries. We introduce the notion of the impact of an edit, measured by the number of times the edited version is viewed. Using several datasets, including recent logs of all article views, we show that an overwhelming majority of the viewed words were written by frequent editors and that this majority is increasing. Similarly, using the same impact measure, we show that the probability of a typical article view being damaged is small but increasing, and we present empirically grounded classes of damage. Finally, we make policy recommendations for Wikipedia and other wikis in light of these findings.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work, web-based interaction

General Terms

Human factors

Keywords

Wiki, Wikipedia, collaboration, vandalism, damage

1. INTRODUCTION

Wikipedia is the great success story of collective action on the Web. It is also a source of wonder: its essential idea – that a useful encyclopedia can be created by allowing *anyone* to create and edit articles – seems absurd. Some people are ignorant, some are malicious, and some are just bad writers. Yet, Wikipedia seems to work. As of this writing, it contains nearly two million articles and ranks among the top ten most visited sites on the Web. Further, some research has found that the accuracy of Wikipedia and *Encyclopaedia Britannica* are roughly equivalent [9].

How does Wikipedia work? As a wiki, every one of its articles can be edited by anyone – there is no credential checking. Changes are visible to everyone immediately, without any review cycle. However, there is intense, ongoing review of articles. Wikipedia has attracted a community of deeply commited editors: for example, Kittur et al. found that by mid-2006 there were hundreds of users who had made over 10,000 edits and well over 10,000 editors who had made more than 100 edits [13]. Editing is made easier by various mechanisms: a *recent changes* page and IRC channel show every edit made to every article, *watch lists* help users monitor articles they care about, and *version histories* help editors quickly roll back objectionable changes.

That Wikipedia works may be amazing, but *how* it works is a research question, one that previous work has addressed in various forms. We build on prior work by developing a new approach to estimating the value of Wikipedia, based on how many people are affected by a change to an article. We pose three specific research questions:

1. **Creating value**: Who contributes Wikipedia's value? Is it the handful of people who edit thousands of times, or is it the thousands of people who edit a handful of times?

2. **Impact of damage**: What is the impact of damage such as nonsensical, offensive, or false content? How quickly is it repaired, and how much of it persists long enough to confuse, offend, or mislead readers?

3. **Types of damage**: What types of damage occur, and how often?

Regardless of the value of Wikipedia itself, these questions matter, as do our results. Many other popular online communities produce artifacts of lasting value [7] through the collective action of their users. Sites such as slashdot.org, reddit.com, and digg.com all rely on user opinions to filter and order the vast volume of online news stories. Users of freedb.org (CDs) and imdb.com (movies) have created large databases that provide value to anyone interested in those topics. These examples, and others, show that the issues of who creates value, and the types and impacts of damaging behavior, are of general interest.

The rest of the paper is organized as follows. First, we survey related work, showing how ours builds upon and extends it. Most notably, our work is the first to compute *value* of edits and the *impact* of damage in terms of how many user views they receive. Estimating this is hard, so

GROUP'07, November 4–7, 2007, Sanibel Island, Florida, USA.

we detail how it is done. We then describe the methods and present the results for each of our three research questions and close with a brief summary.

2. RELATED WORK

The past few decades have seen the emergence of numerous Internet-based social media, including email lists, Usenet, MUDs and MOOs, chat, blogs, wikis, and social networking systems. Researchers have taken advantage of the opportunities created by the large user bases these media have attracted, creating a wide variety of advanced software (e.g., visualization and analysis tools [19, 16], social agents [11], and social navigation aids [23]) and conducting a broad range of empirical research (e.g., on conversational patterns [3], social interaction [5], gender [10], and group-wide patterns [24]).

2.1 Content and quality in social media

The areas we address – who contributes value, and how does a community maintain itself against antisocial behavior – have been widely researched in different social media.

Creating Value. It has been observed widely that a small minority of participants in an online group produce most of the content, while the vast majority of users produce little or no content. The distribution of activity typically takes the form of a *power law*. Whittaker et al. [24] observed this for Usenet postings and Marks [15] for blog links.

Antisocial behavior (damage). Much work has been devoted to characterizing, detecting, and managing behavior like *flaming* (personal attacks) and *spam* (irrelevant content designed for financial or other gain). For example, Slashdot uses a socially-based moderation system that is effective in making it easy to ignore uninteresting or offensive content [14]. More generally, Cosley et al. [7] presented a model and empirical results suggesting that a policy of reviewing content *after* it is "published" (as in Wikipedia) eventually results in quality equivalent to that obtained under a pre-review policy (as in traditional peer-reviewed journals).

2.2 Research on Wikipedia

As Wikipedia has grown, it has attracted research on a variety of issues, including predicting article quality [17], comparing article topics to those in traditional encyclopedias [8], contrasting participation in Wikipedia and in open source software projects [17], and analyzing the development of contributors from an activity-theoretic perspective [4].

2.2.1 Who produces the content?

The issue of who (i.e., what types of editors) contributes Wikipedia's content is a matter of some dispute. Jimmy Wales, one of the founders of Wikipedia, has stated that "2% of the users do 75% of the work" [22], while Swartz [18] has argued that the work is more distributed. Voss [21] provided data on this question by counting number of edits: unsurprisingly, the data showed a power law distribution.

Kittur et al. [13] analyzed the amount of content contributed by different classes of editors, finding that elite users (10,000 or more edits) accounted for over 50% of edits in 2002 but only 20% by mid-2006, due to increasing participation by users with less than 100 edits. Furthermore, Kittur and his colleagues explored whether elite and low-edit groups accounted for different amounts of content. By measuring the total number of words added and deleted in edits, they found that elite editors accounted for a higher proportion of content changes (around 30%) and were more likely than low-edit users to add (rather than delete) words.

Adler and Alfaro [1] developed a reputation system for Wikipedia editors. Their reputation metric futher developed the notion of measuring editors' value contributions by accounting for whether changes introduced by an edit *persisted* over time. Defining *short-lived edits* and *short-lived text* as changes that were at least 80% undone (by edit distance) within a few subsequent edits, they showed that these metrics could be used to compute reputations for authors, and that these reputations predicted well whether an author's text would persist.

We share the concern of these efforts to understand who produces Wikipedia's valuable content. Kittur et al. took an important step by examining content within edits, and Adler and Alfaro took another by considering whether an edit's changes persist. Our work, however, produces two significant advances.

First, we use a more general notion of persistence than Adler and Alfaro, measuring how words persist over time rather than just detecting short-lived changes. Second, we compute how much each word is *viewed* over time. There is no real value in content that no one views, even if there is a lot of it; conversely, content that is viewed frequently has high value, regardless of how much of it there is. Thus, our metric matches the notion of the value of content in Wikipedia better than previous metrics.

2.2.2 Damage in Wikipedia

Over the past few years, the issue of vandalism in Wikipedia, e.g. deliberate and malicious editing of a destructive nature, has received much attention. There have been a number of high-profile cases of vandalism in Wikipedia. For example, Adam Curry allegedly altered pages about the history of podcasting in order to promote his role and diminish that of others [30], Jeffrey Seigenthaler's article stated falsely for months that he was involved in the John F. Kennedy assassination [31], and the comedian Stephen Colbert has even conducted humorous tutorials on how to vandalize Wikipedia [33].

In 2004, Viégas et al. published a seminal paper [20] that popularized the study of Wikipedia in the HCI and CSCW fields, introducing a novel visualization technique for exploring the history of edits to Wikipedia articles. Most vividly, they used this to identify cases of conflict and vandalism. Focusing on one particular class of vandalism, *mass delete*, where all or nearly all of an article's content is deleted, they found that this vandalism was repaired in a median time of 2.8 minutes.

In more recent work, Kittur et al. [12] investigated the occurrence of the key mechanism for repairing vandalism, the revert, and found that its use is growing.

Our work significantly extends that of Viégas. First, we systematically categorize types of damage in Wikipedia and provide estimates on how common they are. Second, building on Kittur et al.'s use of reverts, we propose and evaluate a new vandalism-detecting metric and use it to analyze three orders of magnitude more instances of vandalism than Viégas. Third, we analyze not only how long articles remained in a damaged state, but also how many times they were viewed while in this state. Finally, we use a larger, current, and comprehensive corpus: nearly three more years of

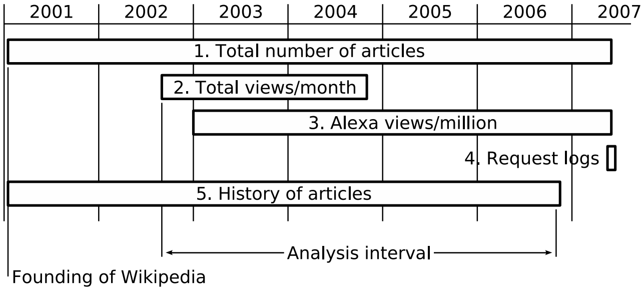

Figure 1: Datasets used in this paper. We consider the period September 1, 2002 to October 31, 2006.

Wikipedia article history (ten times as many revisions) plus additional datasets to estimate the number of views each revision of an article received. Thus, we can better approximate the *impact* of damage on readers.

3. ESTIMATING ARTICLE VIEWS

We define a few critical terms as follows. The *act* of making and saving changes to an article is an **edit**, and the history of an article forms a sequence of content states called **revisions** – i.e., edits are transitions between revisions. Further, there is a special kind of edit, called a **revert**: reverting an article means restoring its content to some previous revision, removing the effects of intervening edits.

3.1 Why measure views?

We measure the value of contributions or the impact of damage in terms of number of **views**. Importantly, this information is required for every point in time during the period of Wikipedia's history under study. It's not enough to know how many views an article receives "now". Instead, we need to know how many views it received (for example) between 9:17 and 9:52 on July 8, 2005 – perhaps because it was in a damaged state.

There is a reason past analyses haven't used view data: it is not available, because the relevant logs no longer exist. However, we have access to several datasets that let us estimate view data.[1]

A word of caution. We assume that one serving of an article by a Wikipedia server is a reasonable proxy for one view of that article by a user. While a human may request an article but not read all of it, or web caching schemes may cause one Wikipedia serving of an article to correspond to many user views of that article, we believe that these factors do not materially affect the accuracy of our view estimates, in particular since we are most interested in comparisons between articles.

3.2 Data Sets

We use five datasets in this work, illustrated in Figure 1. The first four are used to estimate article views, while the fifth is used to estimate the secondary metrics that Sections 4 and 5 focus on. It is included here for completeness.

1. **Total number of articles** in Wikipedia over time [32] is provided by Wikipedia itself.
2. **Total views per month** is also provided by Wikipedia [27], but only for the period August 2002 to October 2004.
3. **Alexa views per million** over time is compiled by Wikipedia [29] from data recorded by Alexa based on use of its Alexa Toolbar browser plugin, a "search and navigation companion" [2]. These data measure, of each million page requests made by users with the plugin installed, how many requests were for pages under wikipedia.org, including all languages. Users self-select to install the plugin, but Alexa data are among the best available estimates of Web page views.
4. **Request logs** from Wikipedia server activity. The Wikimedia Foundation recently provided us access to a sampled log of all requests served by its web servers and internal caching system. This includes all Wikipedia languages plus several other wiki projects; our current analyses make use only of the English language Wikipedia data. The logs contain the timestamp and URL of every 10th HTTP request.[2]

   Our analyses consider log data between April 12 and May 11, 2007. During an average day in this time period, Wikimedia served 100 million English Wikipedia article requests (and a total of 1.7 billion HTTP requests). Even these 10% sampled logs are huge, comprising 10-15 GB of data per day.
5. **History of articles**. Wikipedia provides a historical archive of its content, i.e. the text of all articles with complete edit history. We analyzed the 1.2 TB archive containing changes through the end of October 2006.

A tempting proxy for article views is article edits. However, we found essentially no correlation between views and edits in the request logs. Therefore, we must turn to a more elaborate estimation procedure.

3.3 Computing Article Views

We need to compute the number of article views during a specific interval – how many times was article X viewed during the interval t_1 to t_2? We do this by computing the article's *view rate* – e.g., how many views per day did article X have at time t – from which it is straightforward to compute views. Specifically, we compute:

$$r(X,t) = r(X,\text{now}) \times \frac{R(t)}{R(\text{now})} \times \frac{Z(\text{now})}{Z(t)}$$

where $\boldsymbol{r(X,t)}$ is the view rate of article X at time t, $\boldsymbol{R(t)}$ is the view rate of Wikipedia as a whole at time t (i.e., $R(t) = \sum_a r(a,t)$ for each article a), and $\boldsymbol{Z(t)}$ is the number of articles in Wikipedia at time t.

Intuitively, to compute the historical view rate of an article, we take its current view rate, *shrink* it to compensate for shrinkage of Wikipedia's view rate as a whole, and then *expand* it to compensate for the smaller number of articles. This computation is based on the assumption that the view rates of articles, relative to each other, are fairly stable.

The values of the five terms in the above formula are themselves computed as follows. $\boldsymbol{r(X,\text{now})}$ and $\boldsymbol{R(\text{now})}$ are

[1] Our estimation tool is available online at http://www.cs.umn.edu/~reid/views.tar.bz2.

[2] 1.5% of the log data were lost between Wikimedia and us due to dropped packets and interruptions of our collector process. Our estimates compensate for this small loss.

average values from the request logs (data set 4). "Now" is the center of the request log data period, i.e. April 24, 2007, while $\boldsymbol{Z(t)}$ and $\boldsymbol{Z(\mathbf{now})}$ are interpolated from the total number of articles (data set 1).

$\boldsymbol{R(t)}$ is the most complex to compute. If t falls within the period covered by the total views-per-month data from Wikipedia (data set 2), it is interpolated from those data. If not, we interpolate it from *scaled* Alexa data (data set 3). Intuitively, what we are after is a scaling of the Alexa data so that it matches well both the old views-per-month data and the new $R(\text{now})$ computed from request logs. This is done by taking the linear regressions of the log of the Alexa data and the log of the views-per-month data during the period of overlap, then scaling up the Alexa data until the linear regressions intersect at the center of this period. This also results in a close match between the scaled Alexa data and $R(\text{now})$: 97 and 104 million views per day, respectively. This scaled Alexa data is what we use to interpolate $R(t)$.[3]

If an article had aliases at time t, $r(a, t)$ is computed for each alias a as well, and the final view rate is the sum of the view rates of the article and each of its aliases. Historical view rates for articles or aliases which no longer exist will be zero, but we believe this does not materially affect our results because few articles and aliases are deleted, and those that are are obscure.

For example, suppose that we wish to calculate the number of views of article Y, which has no aliases, from June 9 to June 14, 2004. Suppose that $r(Y, \text{now})$ is 1,000 views per day, and recall that $R(\text{now})$ is 104 million views per day. We first compute the views per day of article Y on June 12, the center of the period.

1. $Z(\text{now}) = 1{,}752{,}524$, interpolated from the values for April 16 (1,740,243) and May 1, 2007 (1,763,270).

2. $Z(\langle\text{June 12, 2004}\rangle) = 284{,}240$, interpolated from the values for May 13 (264,854) and July 10 (302,333).

3. $R(\langle\text{June 12, 2004}\rangle) = 5{,}019{,}355$, interpolated from the Wikipedia views-per-month data of 2,400,000 views per day on May 15 and 5,300,000 on June 15.

Applying the formula, $r(Y, \langle\text{August 12, 2004}\rangle) = 298$. Because the period is six days long, we estimate the number of views as six times this value, or 1,785 views. The calculation would proceed similarly for the period November 9 to November 14, 2004, except $R(\langle\text{November 12, 2004}\rangle)$ would be interpolated from scaled Alexa data, because the last Wikipedia views-per-month datum is for October 15, 2004.[4]

4. RQ1: CREATING VALUE

4.1 Methods

4.1.1 Persistent word views

As a proxy for the value contributed by an edit, we use the *persistent word view* (PWV), the number of times any given word introduced by an edit is viewed. PWV builds on the notion of an article view: each time an article is viewed, each of its words is also viewed. When a word written by editor X is viewed, he or she is credited with one PWV.

#	Editor	Article text
1	Carol	alpha bravo charlie delta
2	Denise	alpha alpha bravo delta charlie
3	Bob	alpha bravo charlie echo delta
4	Bob	alpha bravo echo foxtrot delta
5	Alice	alpha delta echo foxtrot

Figure 2: Example revision history.

Two key insights drive this metric. First, authors who write content that is read often are empirically providing value to the community. Second, if a contribution is viewed many times without being changed or deleted, it is likely to be a valuable. Of course, this metric is not perfect: the concept of value is dependent on the needs and mental state of the reader. One might imagine a single fact, expressed in only a few words, that provides enormous value to the one reader who really needs that fact. In a large pseudonymous reading community like Wikipedia, capturing a notion of value that depends on the specific information needs of the readers is outside the scope of this work.

For example, see Figure 2. Assuming that each page is viewed 100 times after each edit, Carol has accrued at least 1,200 PWVs: 400 from *bravo* (because she wrote it and it was present for 4 edits), 300 from *charlie*, and 500 from *delta* (even though it was moved several times).

The case of *alpha* is problematic because it is ambiguous. When Bob deleted an *alpha*, whose did he delete: Carol's or Denise's? Words carry no identifier, so it is impossible to tell for certain without understanding the text of the two edits. In cases of ambiguity, we choose randomly. Carol could have 1,300 or 1,700 PWVs depending on whether or not her *alpha* was chosen. Over the trillions of PWVs analyzed, these random choices have little effect on the results.

4.1.2 Calculating PWVs

PWVs are calculated per-article, and the final score for each editor is the sum of his or her scores over all articles. The "owner" of each PWV is determined by comparing the text of subsequent article edits, data contained in the history of articles (data set 5 above); specifically, we:

1. Remove punctuation (except hyphens) and wiki markup from the texts of the old and new edits.

2. Eliminate letter case.

3. Remove stopwords, because very common words carry little information. We use the same word list used by Wikipedia until it began using the Lucene full-text search engine [25].

4. Sort each list of words, because we analyze the appearance and disappearance of words, not their movement within an article.

5. Compare the new and old word sequences to determine which words have been added and which deleted.

6. The editor who made the new edit begins accruing PWV credit for added words, and editor(s) who wrote the deleted words stop accruing PWV credit for them.

[3] The time period between Wikipedia's founding and the beginning of the views-per-month data is excluded from analysis in the present work.

[4] In actuality, these computations are done in floating-point seconds, but we simplify here for clarity.

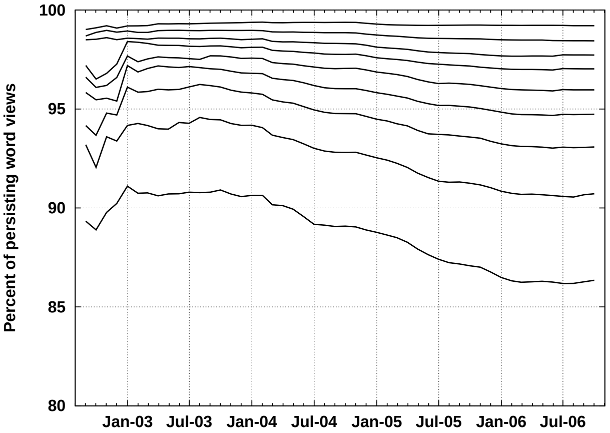

Figure 3: Percentage of PWVs according to the decile of the editor who wrote them.

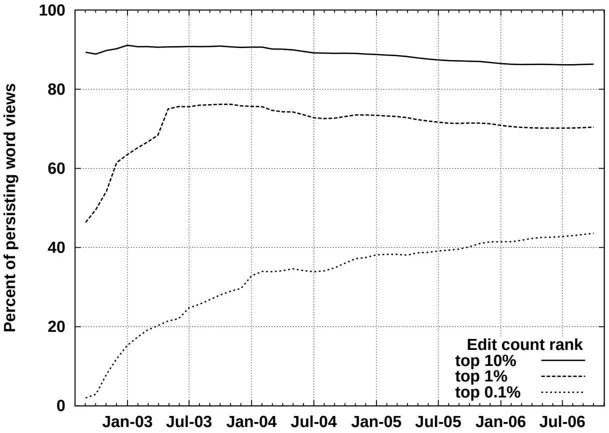

Figure 4: PWV contributions of elite editors.

Our software does not track persistent words if text is "cut-and-pasted" from one article to another. If an editor moves a block of text from one article to another, PWVs after the move will be credited to the moving editor, not to the original editors. This problem is challenging, because edits are per-article, making it difficult to detect where the text moved to, or even if it moved to only one place.

An editor can work either anonymously, causing edits to be associated with the IP address of his or her computer, or while logged in to a pseudonymous user account, causing edits to be associated with that pseudonym. We exclude anonymous editors from some analyses, because IPs are not stable: multiple edits by the same human might be recorded under different IPs, and multiple humans can share an IP.

4.1.3 Dealing with Reverts

Editors who revert do not earn PWV credit for the words that they restore, because they are not adding value, only restoring it; rather, the editors whose words they restore regain credit for those words.

Reverts take two forms: *identity revert*, where the post-revert revision is identical to a previous revision, and *effective revert*, where the effects of prior edits are removed (perhaps only partially), but the new text is not identical to any prior revision. Identity reverts are unusually common, because Wikipedia includes a special mechanism through which any editor can easily revert a page to a previous edit, and because the official Wikipedia guide to resolving vandalism recommends using this mechanism. Kittur et al. [12] report that of identity reverts and effective reverts which could be identified by examining edit comments, 94% are identity reverts. There are probably other effective reverts, because some editors do not clearly label their edits, but detecting these is challenging because it requires understanding the *intent* of the editor. In this paper, we consider only identity reverts.

4.2 Results

We analyzed 4.2 million editors and 58 million edits. The total number of persistent word views was 34 trillion; or, excluding anonymous editors, 25 trillion. 300 billion PWVs were due to edits before the start of our analysis and were excluded. 330 billion PWVs were due to bots – autonomous or semi-autonomous programs that edit Wikipedia.[5][6]

Figure 3 shows the relative PWV contributions of editors divided by edit count decile. From January 2003 to February 2004, the 10% of editors with the most edits contributed about 91% of the PWVs. Then, until February 2006, Wikipedia slowly became more egalitarian, but around February 2006, the top 10% re-stabilized at about 86% of PWVs. Growth of PWV share increases super-exponentially by edit count rank; in other words, elite editors (those who edit the most times) account for *more* value than they would given a power-law relationship. Figure 4 zooms in; editors with the top 0.1% of edits (about 4,200 users) have contributed over 40% of Wikipedia's value. Collectively, the ten editors with the most PWVs contributed 2.6% of all the PWVs.

4.3 Discussion

Editors who edit many times dominate what people see when they visit Wikipedia. The top 10% of editors by number of edits contributed 86% of the PWVs, and top 0.1% contributed 44% – nearly half! The domination of these very top contributors is increasing over time.

Of the top 10 contributors of PWVs, nine had made well over 10,000 edits. However, only three of these users were also in the top 50 ranked by number of edits. The number one PWV contributor, Maveric149, contributed 0.5% of all PWVs, having edited 41,000 times on 18,000 articles. Among the top PWV contributors, WhisperToMe (#8) is highest ranked by number of edits: he is #13 on that list, having edited 74,000 times on 27,000 articles.

Exploring the list of top editors by edit count, we notice something interesting: the list is filled with bots. They occupy the top 4 slots, 9 of the top 10, and at least 20 of the top 50. One the other hand, the list of top editors by PWV

[5] We identified bots by taking the union of (a) editors with usernames ending in "bot", followed by an optional digit, that had at least 100 edits and (b) users listed on Wikipedia's list of approved bots [28].

[6] Some damage-reverting bots had a bug causing a few reverts to become non-identity reverts. Because our software could not detect these reverts, it treated the situations as removal of all text and replacement with entirely new text. Effectively, these bots "stole" PWVs from their rightful owners. Our measurements show that these bugs resulted in only about 0.5% of PWVs being stolen.

is filled with humans: only 2 bots appear in the top 50, and none in the top 10. This suggests, perhaps reassuringly, that people still matter.

5. RQ2: IMPACT OF DAMAGE

5.1 Methods

5.1.1 Damaged article views

We interpret damage differently from value, because in some cases only a few changed words can result in a worse-than-useless article. For instance, adding or removing the word "not" or changing a key date can mislead a reader dangerously. Therefore, we classify a revision simply as "damaged" or "not damaged" rather than counting the number of damaged words. The key metric in this analysis is the *damaged article view* (DAV), which measures the number of times an article was viewed while damaged.

5.1.2 Calculating DAVs

We use the opinions of the Wikipedia community in deciding which revisions are in a damaged state. Note that we focus on revisions that are damaged, *not* the edits which repair the damage, because these are the revisions which lead to damaged article views. Revisions that are subsequently reverted (using an identity revert) are considered as candidates for the "damaged" label. To distinguish damage repair from disagreement or other non-damaging behavior, we look for edit comments on the reverts that suggest intent to repair damage.

We assume that for the purpose of damage analysis, most incidents of damage are repaired by identity reverts. This assumption is motivated by two factors. First instructions on Wikipedia itself have, since the beginning of our analysis period, recommended the use of identity reverts for damage repair [26]. Second, repairing damage using the wiki interface to make an identity revert is easier than manually editing away the damage.

It is important to note that our method is not foolproof, as editors sometimes make mistakes or place overheated rhetoric into the comments of reverts, labeling each other vandals when a neutral reader would consider the situation simply a content dispute. We also cannot discover damage which was not yet repaired by the end of the article history. Nonetheless, as our results below show, our method is essentially sound.

At any given instant, a revision is in zero or more of the following states. State membership is determined by how future editors react to the revision, so it can only be determined in retrospect. The edit comments and timestamps necessary for these computations are available in the history of articles (data set 5). Figure 5 shows the relationships of the various states informally.

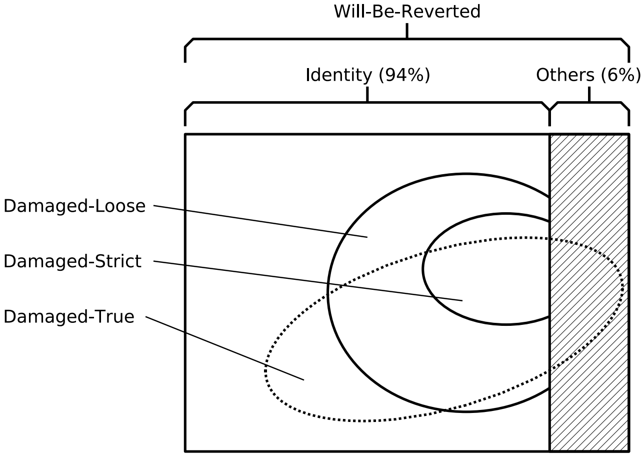

Figure 5: Classes of revisions: Damaged-Truth is the conjectured true set of damaged revisions, while Damaged-Loose and Damaged-Strict are increasingly restrictive subsets of Will-Be-Reverted designed to approximate Damaged-True.

#	Time	MD5	Editor	Edit comment
1	9:19	24bd	Carol	new article
2	10:04	6f59	Denise	clarify
3	10:19	2370	Bob	arrrrr!!!
4	10:37	02ac	Bob	shiver me timbers!!!
5	10:56	6f59	Alice	revert vandalism

Figure 6: Example revision history.

- **Will-Be-Reverted** (WBR): a revision which will reverted by a future edit. Several revisions in a row might be reverted by the same revert; we refer to such a group of revisions as a *WBR sequence*. We detect reverts by comparing the MD5 checksums of the texts of each revision.
- **Damaged-Loose** (D-Loose): a WBR revision where the future revert's edit comment suggests either (a) explicit intent to repair vandalism *or* (b) use of revert-helper tools or autonomous anti-vandalism bot activity. Specifically, the revert's edit comment matches a regular expression for criterion (a) or another for (b).[7]
- **Damaged-Strict** (D-Strict): a D-Loose revision that matches criterion (a). This more-selective state is intended to trade some of the recall of D-Loose for greater precision.
- **Damaged-True** (D-True): a revision that is damaged. The damage may have appeared in this revision, or it may persist from a prior one.

For example, consider the revision history in Figure 6. Revision 5 reverts back to revision 2 – we know this because the MD5 checksums are the same – discarding the effects of revisions 3 and 4. Therefore, revisions 3 and 4 are in state WBR; additionally, because Revision 5's edit comment matches the criteria for D-Strict, these two revisions are also in D-Loose and D-Strict. Finally, if each revision were viewed 10 times, there would be 20 DAVs generated by this sequence.

Both D-Loose and D-Strict limit the distance between the first damaged revision and the repairing revert to 15 revisions. This is to avoid false positives due to a form of damage where someone reverts an article to a long-obsolete revision and then marks this (damaging) revert as vandalism repair.

[7] Our classification software, which includes these expressions, is available by emailing the authors.

We believe it is reasonable to assume that essentially all damage is repaired within 15 revisions.

The purpose of states D-Loose and D-Strict is to be proxies for the difficult-to-determine state D-True. To evaluate the two metrics for this purpose, three human judges independently classified 676 WBR revisions, in 493 WBR sequences selected randomly from all WBR sequences. Classification included a best effort to figure out what was going on, which often included a minute or two of research to verify information or clarify unfamiliar topics, words, or links. The edit comment of the final revert was hidden in order to avoid biasing the judges.

Judges classified revisions into the following three classes:

- **Vandalized-Human** (V-Human): WBR revisions that introduce or persist clearly deliberate damage. We attempted to follow the Wikipedia community definition of vandalism, which emphasizes *intent*.

- **Damaged-Human** (D-Human): WBR revisions which introduce or persist damage (a superset of V-Human).

- **Other**: All other WBR revisions. Frequent examples were content disputes or editors changing their minds and reverting their own work.

Determining whether a revision was V-Human or just D-Human is difficult because it requires assessing the intent of the editor. Indeed, despite written guidelines and calibration by judging together a smaller independent set of WBR revisions, there were considerable differences between the judges. From the reader's perspective, the intent behind damage is irrelevant, so we consider further only D-Human.

We use these judgements to evalute the effectiveness of the classes D-Loose and D-Strict as proxies for D-True. Of the 676 revisions judged, all three judges agreed on 437 (60%), while the class of the remaining 239 (35%) was determined by 2-1 majority. We assumed that revisions judged D-Human by a majority of judges, and no others, were in class D-True.

By this measure, 403 revisions (60%) were in D-True. The automatic D-Strict classifier had a precision of 0.80 but a recall of only 0.17, i.e., within the judged revisions, 80% of D-Strict revisions were in D-True, but only 17% of D-True revisions were in D-Strict; D-Strict is therefore not a reasonable proxy for D-True.

On the other hand, the precision and recall of D-Loose were 0.77 and 0.62 respectively. Clearly, D-Loose suffers from both false negatives and false positives. The former arise when editors revert damage but do not label their actions clearly, while the latter can be seen in content disputes, as described previously. While imperfect, D-Loose is a reasonable proxy for D-True. The remainder of this section will consider D-Loose only.

5.2 Results

We found 2,100,828 damage incidents (i.e., D-Loose sequences). 1,294 overlapped the end of our study period, so there were 2,099,534 damage-repair reverts. No incidents overlapped the beginning of the study period. These incidents comprised 2,955,698 damaged revisions, i.e. an average sequence comprised 1.4 damaged revisions before repair. The study period contained 57,601,644 revisions overall, so about 5% of revisions were damaged.

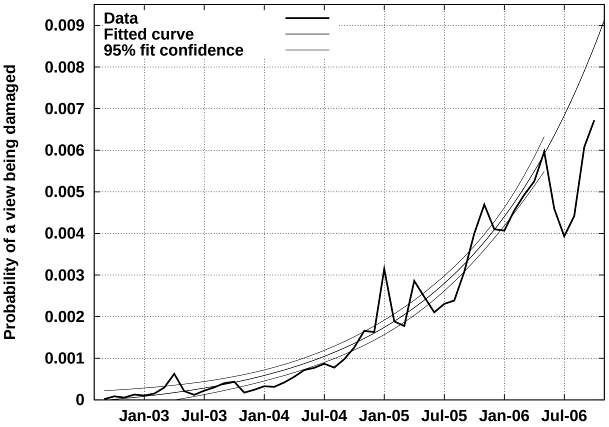

Figure 7: Probability of a typical view returning a damaged article. Both our data and a fitted exponential curve (with boundary lines representing the 95% confidence interval) are shown. We fitted through June 2006, when widespread use of autonomous vandalism-repair bots began.

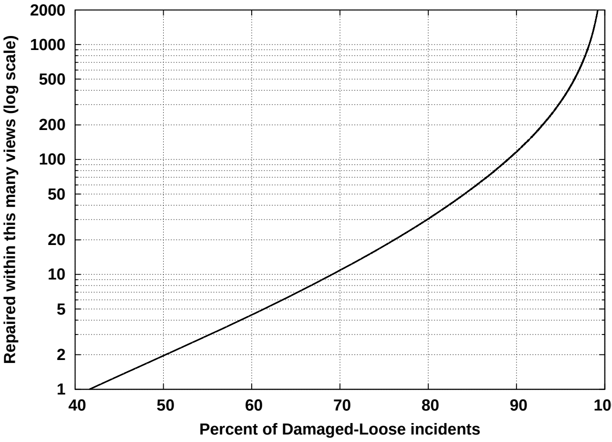

Figure 8: Rapidity of damage repair: 42% of damage incidents are repaired within one estimated view, meaning that they have essentially no impact.

During the study period, we estimate that Wikipedia had 51 billion total views. Of these, 188 million were damaged – 139 million by anonymous users – meaning that the overall probability of a typical view encountering damage was 0.0037. In October 2006, the last month we analyzed, it was 0.0067. Figure 7 illustrates the growth of this probability over time. In particular, the data through June 2006 fits the exponential curve $e^{0.70x-8.2} - 0.0003$.

Figure 8 illustrates the rapidity of damage repair. 42% of damage incidents are repaired essentially immediately (i.e., within one estimated view). This result is roughly consistent with the work of Viégas et al. [20], which showed that the median persistence of certain types of damage was 2.8 minutes. However, 11% of incidents persist beyond 100 views, 0.75% – 15,756 incidents – beyond 1000 views, and 0.06% – 1,260 incidents – beyond 10,000 views. There were 9 outliers beyond 100,000 views and 2 beyond 500,000; of these, 8 were false positives (the other was the “Message” incident discussed below). The persistence of individual incidents of damage has been relatively stable since 2004, so the increasing probability of damaged views indicates a higher rate of damage.

A possible cause for this increase is that users may be writing edit comments differently, increasing the number of edit comments that fit the D-Loose pattern. To test this hypothesis, we judged the precision and recall of D-Loose for a sample of 100 WBR sequences (containing 115 revisions) from 2003 and earlier, using threee judges and majority opinion as above. We found that the recall of D-Loose over this sample was 0.49, compared to 0.64 for a sample of 100 sequences (134 revisions) from 2006. Thus, commenting behavior has changed, and this change explains about one-third of the increase in probability of a damaged view. Damage was only 14 times more impactful in 2006 than 2003, not 18 times.

5.3 Discussion

While the overall impact of damage in Wikipedia is low, it is rising. The appearance of vandalism-repair bots in early 2006 seems to have halted the exponential growth (note the dramatic drop in view probability after June 2006), but it is too early to tell what the lasting impact will be.

Many of the editors who performed the reverts we analyzed appear to have treated vandalism and other damage the same way. For instance, it is common to find edit comments asserting vandalism repair on reverts of revisions which are apparently users practicing editing on a regular article, something which is not welcome but is (by policy) explicitly not vandalism. This makes sense, as from the reader's perspective, damage is damage regardless of intent.

The most viewed instance of damage was deletion of the entire text of the article "Message", which lasted for 35 hours from January 30 to February 1, 2006 and was viewed 120,000 times. The popularity of this article is partly explained by the fact that it is (at the time of this writing) the top Google search result for "message". It may seem odd that such a high traffic article was not fixed more quickly. However, it is not edited very frequently, only about once every 19 days. (This reminds us that there is no correlation between view rate and edit rate.) Put another way, the tens of thousands of people who viewed the article while it was damaged simply may not have included any Wikipedia editors. Further, maybe this type of damage is not as inviting of a fix as others, such as obscenities.

One example of a high-traffic article which suffered greatly from damage was "Wiki" in October 2006. Of the 3.1 million estimated views during that month, 330,000 were damaged, over 10%. This page was bombarded with damaging edits having no apparent pattern, most of which were repaired rapidly within minutes or hours but which had dramatic impact due to their sheer numbers. Another interesting example is "Daniel Baldwin," an article about an American actor, damaged by a single incident which lasted for three months. In October 2005, someone vandalized the article by deleting the introduction and replacing it with with text explaining (incorrectly) that Baldwin was a college student and frequent liar. The next day, someone removed the bogus text but failed to restore the introduction; this situation persisted through five more revisions until someone finally made a complete repair in February 2006.

We have two policy suggestions for combating damage, both based on distributing repair work to humans. The first is to ensure that revisions are reviewed by n humans within a few seconds of being saved. The second is to ensure that each article is on at least n users' watch lists. Assuming an edit rate of 280,000 edits per day (the average rate we observed in our log analysis), and assuming it takes 30 seconds to determine if an average revision is damaged, schemes like these would require about $n \times 28{,}000$ reviewers averaging five minutes of daily work.

6. RQ3: TYPES OF DAMAGE

We have mentioned that for each damaged edit, the level of damage depends on what exactly the damage is: for example, a reader might consider deleting all of an article's content more damaging than adding nonsense, and false information more damaging still. Consequently, to understand the impact of different damages, it is meaningful to define different types of damage from the reader's perspective and provide estimates of how often each type of damage occurs.

6.1 Methods

Based on the experience of judging edits for RQ2 and developing the tools for doing so, we present a list of features exhibited by Wikipedia damage, aiming for comprehensiveness. These features, with comparisons to the anecdotal categories of Viégas et al. [20], are as follows:

- **Misinformation**: Information which is false, such as changed dates, inappropriate insertion of "not", or stating incorrectly that a public figure is dead. (No analogue in Viégas.)

- **Mass delete**: Removal of all or nearly all of an article's content. (Same as Viégas.)

- **Partial delete**: Removal of some of an article's content, from a few sentences to many paragraphs. (No analogue in Viégas.)

- **Offensive**: Text offensive to many users, such as obscenities, hate speech, attacks on public figures, etc. This is a broad category, ranging e.g. from simple "you suck" to unexpected pornography. (Includes Viégas' *offensive copy.*)

- **Spam**: Advertisements or non-useful links. (No analogue in Viégas.)

- **Nonsense**: Text that is meaningless to the reader, for example "Kilroy was here", characters that do not form words, obviously unrelated text, and technical markup leaking into formatted pages. (Includes Viégas' *phony copy.*)

- **Other**: Damage not covered by the other six types.

Viégas' *phony redirection* is not included above because we observed only one instance (and it was better described as Offensive), and we believe that *idiosyncratic copy* ("text that is clearly one-sided, not of general interest, or inflammatory") better describes disputed content, not damage.

In reflecting on the results for RQ2, we observed that most D-Human sequences consisted of one or more related edits that formed a coherent single incident. Therefore, to focus the effort of our judges, we used the D-Human sequence rather than individual revisions as the unit of analysis. Of

Feature	%	Agreement			Reliability	
		3v0	2v1	1v2	PF	Ja
Nonsense	53	108	56	70	0.66	0.46
Offensive	28	57	30	29	0.66	0.49
Misinformation	20	28	34	64	0.45	0.22
Partial Delete	14	35	7	20	0.83	0.56
Spam	9	25	3	6	0.89	0.74
Mass Delete	9	23	5	3	0.82	0.74
Other	5	1	15	21	0.06	0.27

Figure 9: Distribution of damage features. % is the percentage of D-Human sequences where the feature applies (determined by majority vote), while the *Agreement* columns list how many times all (*3v0*), two of the three (*2v1*), and only one of the judges (*1v2*) believed the feature applied. (Percentages do not sum to 100 because features are not mutually exclusive.) *PF* (*proportion full*) gives the proportion assigned unanimously (i.e. $PF = 3v0/(3v0 + 2v1)$), while *Ja* gives the Jacquard statistic: the number of times all judges assigned the feature divided by the number of times any assigned the feature, i.e. $Ja = 3v0/(3v0 + 2v1 + 1v2)$.

the 493 WBR sequences analyzed in RQ2, 308 were classified as D-Human. These 308 sequences form the basis of this section's analysis.

After calibration on a different sample of D-Human edit sequences, three judges independently classified the sequences, applying as many of the damage features as were appropriate. As in RQ2 above, we used a "majority vote" procedure, i.e. a feature applied to a sequence if at least two of the three judges believe that it does.

6.2 Results

Figure 9 summarizes our results. It is not surprising that agreement was highest for Spam, Mass Delete, and Partial Delete, since these features do not require much judgement. On the other hand, what's offensive or nonsense is somewhat subjective, and misinformation may be subtle. Finally, the low number of D-Human sequences labeled Other indicate that our categories are relatively comprehensive.

6.3 Discussion

From the perspective of Wikipedia, all damage is serious because it affects the credibility of Wikipedia's content. However, there are specific factors that we can use to assess more precisely the implications of our results. First, how *common* is a given type of damage? If a particular type is infrequent, we need not worry as much. Second, what is the potential *impact* on readers? If there is little harm, we need not worry as much even if occurence is frequent. Finally, how easy is it to *detect* automatically? Even if damage is not automatically repaired, automatic notification of human editors can speed repair.

With this in mind, Mass Delete and Nonsense are low-impact types of damage. The former is relatively uncommon and trivial to detect automatically. The latter, while common, damages only presentation, not content, except in cases where its sheer bulk overwhelms content. For example, one incident consisted of the insertion of thousands of repetitions of a string of Korean characters into the article "Japan". (Interestingly, the characters formed hate speech, but we classified the incident as Nonsense because few readers of the English Wikipedia understand Korean.) Spam and Partial Delete are somewhat higher impact, because they are tricky to detect automatically (useful edits introduce links and remove text all the time); also, Spam wastes readers' time and Partial Delete may cause the omission of important information.

Offensive damage is troublesome because it is common (28% of incidents) and potentially highly impactful – offensive content damages the reputation of Wikipedia and drives away readers. Automatic detection of offensive damage is plausible in some cases (e.g., detecting obscenties) but harder in the general case due to the complexity of offensive speech and the difficulty of analyzing images automatically.

Misinformation may be the most pernicious form of damage. It is both common (20% of incidents) and difficult to detect. Automatic detection is essentially impossible because it requires understanding the content of the page, and people who visit a page are typically there to learn about its topic, not because they understand it well. An intriguing and subtle example is that of the "Uchi-soto" article, which discusses a specific facet of Japanese language and social custom. A (presumably well-meaning) editor changed the translation of the word *uchi* from *inside* to *house* – both are correct, but *inside* is the one appropriate for this article. This error could only be detected by a reader with sophisticated knowledge of Japanese.

7. SUMMARY

Wikipedia *matters*. It is widely used and immensely influential in contemporary discourse. It is the definitive exemplar of collective action on the Web, producing a large, successful resource of great value.

Our work has set the scientific study of Wikipedia – and, by extension, study of other online collective action communities – on a much firmer basis than ever before. Most fundamentally, we offer a better way to measure the phenomena people care about. Others have used author-based measures, counting edits to approximate the value of contributions and measuring repair time to approximate impact of damage. We use reader-based measures. We approximate both the value of contributions and the impact of damage by estimating the number of times they were viewed.

Our view-based metrics let us both sharpen previous results and go beyond them. Others have shown that 1% of Wikipedia editors contributed about half of edits [6]. We show that *1/10th of 1% of editors contributed nearly half of the value*, measured by words read. Others have shown that one type of damage was repaired quickly [20]. We show this for all types of damage. We also show what this result means for readers: 42% of damage is repaired almost immediately, i.e., before it can confuse, offend, or mislead anyone. Nonetheless, there are still hundreds of millions of damaged views. We categorize the types of damage that occured, show how often they occured, describe their potential impact on readers, and discuss how hard (or easy) they are to detect automatically. We give examples of especially impactful damage to illustrate these points. Finally, we show that the probability of encountering damage increased exponentially from January 2003 to June 2006.

What are the implications of our results? First, because a very small proportion of Wikipedia editors account for most of its value, it is important to keep them happy, for example by ensuring that they gain appropriate visibility and status. However, turnover is inevitable in any online community. Wikipedia should also develop policies, tools, and user interfaces to bring in newcomers, teach them community norms, and help them become effective editors.

Second, we speculate that the exponential increase in the probability of encountering damage was stopped by the widespread use of anti-vandalism bots. It is likely that vandals will continue working to defeat the bots, leading to an arms race. Thus, continued work on automatic detection of damage is important. Our results suggest types of damage to focus on; the good news is that the results show little subtlety among most vandals. We also generally believe in augmentation, not automation. That is, we prefer *intelligent task routing* [7] approaches, where automation directs humans to potential damage incidents, but humans make the final decision.

8. ACKNOWLEDGEMENTS

We thank Tim Starling and his colleagues at the Wikimedia Foundation for making request log data available to us and providing technical assistance in its processing. We also thank the members of wikitech-l for their technical help. Aaron Swartz suggested the ideas that developed into Section 4. Finally, we gratefully acknowledge the support and feedback of members of our research group. This work has been supported by the National Science Foundation under grants IIS 03-24851 and IIS 05-34420.

9. REFERENCES

[1] Adler, B. and Alfaro, L. A content-driven repu- tation system for the Wikipedia. In *Proc. WWW*. 2007.

[2] Alexa Internet, Inc. Quick tour. http://www.alexa.com/site/help/quicktour (2007).

[3] Arguello, J. et al. Talk to me: Foundations of successful individual-group interactions in online communities. In *Proc. CHI*. 2006.

[4] Bryant, S. L. et al. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP*. 2005.

[5] Cherny, L. Talk to me: Foundations of successful individual-group interactions in online communities. In *Conversation and Community: Chat in a Virtual World*, Cambridge University Press, 1999.

[6] Chi, E. Long tail of user participation in Wikipedia. http://asc-parc.blogspot.com/2007/05/ (May 2007).

[7] Cosley, D. et al. Using Intelligent Task Routing and Contribution Review to Help Communities Build Artifacts of Lasting Value. In *Proc. CHI*. 2006.

[8] Emigh, W. and Herring, S. C. Collaborative authoring on the Web: A genre analysis of online encyclopedias. In *Proc. HICSS*. 2005.

[9] Giles, J. Internet encyclopedias go head to head. *Nature*, *438* (2005), 900–901.

[10] Herring, S. C. Gender and democracy in computer-mediated communication. In R. Kling, ed., *Computerization and Controversy: Value Conflicts and Social Choices*, Academic Press, 1996. 2nd edition.

[11] Isbell, Jr., C. L. et al. Cobot in LambdaMOO: A social statistics agent. In *Proc. AAAI*, 2000.

[12] Kittur, A. et al. He says, she says: Conflict and coordination in Wikipedia. In *Proc. CHI*. 2007.

[13] Kittur, A. et al. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. alt.CHI*. 2007.

[14] Lampe, C. and Resnick, P. Slash (dot) and Burn: Distributed Moderation in a Large Online Conversation Space. 2004.

[15] Marks, K. Power laws and blogs. http://homepage.mac.com/kevinmarks/powerlaws.html (2003).

[16] Smith, M. A. and Fiore, A. T. Visualization components for persistent conversations. In *Proc. CHI*. 2001.

[17] Stivilia, B. et al. Assessing information quality of a community-based encyclopedia. In *Proc. Information Quality*. 2005.

[18] Swartz, A. Who writes wikipedia? http://www.aaronsw.com/weblog/whowriteswikipedia (September 2006).

[19] Viégas, F. B. and Donath, J. S. Chat circles. In *Proc. CHI*. 1999.

[20] Viégas, F. B. et al. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. CHI*. 2004.

[21] Voss, J. Measuring Wikipedia. In *Proc. Scientometrics and Infometrics*. 2005.

[22] Wales, J. Jimmy Wales talks Wikipedia. http://writingshow.com/?page_id=91 (December 2005).

[23] Wexelblat, A. and Maes, P. Footprints: History-Rich tools for information foraging. In *Proc. CHI*. 1999.

[24] Whittaker, S. et al. The dynamics of mass interaction. In *Proc. CSCW*. 1998.

[25] Wikimedia Foundation. Stop word list. http://meta.wikimedia.org/w/index.php?title=Stop_word_list&oldid=313397 (2006).

[26] Wikipedia. Vandalism in progress. http://en.wikipedia.org/w/index.php?title=Wikipedia:Archive/Wikipedia:Vandalism_in_progress/History&oldid=188844 (August 2002).

[27] Wikipedia. Page requests per day. http://stats.wikimedia.org/EN/TablesUsagePageRequest.htm (October 2004).

[28] Wikipedia. Bots/Status. http://en.wikipedia.org/w/index.php?title=Wikipedia:Bots/Status&oldid=133390929 (May 2006).

[29] Wikipedia. Awareness statistics. http://en.wikipedia.org/w/index.php?title=Wikipedia:Awareness_statistics&oldid=129505430 (May 2007).

[30] Wikipedia. Podcast. http://en.wikipedia.org/w/index.php?title=Podcast&oldid=133330735 (May 2007).

[31] Wikipedia. Seigenthaler Controversy. http://en.wikipedia.org/w/index.php?title=Seigenthaler_controversy&oldid=132296396 (May 2007).

[32] Wikipedia. Size of Wikipedia. http://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=127300877 (May 2007).

[33] Wikipedia. Wikipedia in culture. http://en.wikipedia.org/w/index.php?title=Wikipedia_in_culture&oldid=133473824 (May 2007).