Deletion Discussions in Wikipedia: Decision Factors and Outcomes

Jodi Schneider
Digital Enterprise Research
Institute
National University of Ireland
Galway,Ireland
jodi.schneider@deri.org

Alexandre Passant
Digital Enterprise Research
Institute
National University of Ireland
Galway,Ireland
alexandre.passant@deri.org

Stefan Decker
Digital Enterprise Research
Institute
National University of Ireland
Galway,Ireland
stefan.decker@deri.org

ABSTRACT

Deletion of articles is a common process in Wikipedia, in order to ensure the overall quality of the encyclopedia. Yet, there is a need to better understand the procedures in order to promote the best decisions without unnecessary community work. In this paper, we study deletion in Wikipedia, drawing from factor analysis, and taking an in-depth, content-analysis-based approach. We address three research questions: First, what factors contribute to the decision about whether to delete a given article? Second, when multiple factors are given, what is the relative importance of those factors? Third, what are the outcomes of deletion discussions, both for articles and for the community? We find that multiple factors contribute to the assessment of an article, and we discuss their relative frequency. Further, we show how the assessment timeline focuses attention on improving borderline articles that have the potential to meet Wikipedia's content inclusion policies, and we highlight the role of novice contributors in this improvement process.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organizational Interfaces—*Collaborative computing*

General Terms

Human Factors

Keywords

collaboration and conflict, decision-making, Wikipedia, Articles for Deletion, factors analysis, online argumentation, values, novices

1. INTRODUCTION

Collaborative, voluntary contributions are at the heart of Wikipedia's quality mechanisms [18]. To ensure that the right topics are covered, the encyclopedia relies on procedures, wanted article lists, and willing participants. Thus the knowledge of the active participants, regarding both content and procedures, is key to ensuring

improvement of individual articles and of the encyclopedia as a whole. Recent studies have pointed to a gender gap [24], which the media has suggested may create selection bias regarding which articles are written and developed [10]. Determinations of which articles are retained for further development may also depend on who makes the decisions. Thus, recently, in addition to an ongoing concern for managing the content of the encyclopedia, there has been an increasing interest in managing the *people* behind that content, by attracting and retaining new editors.

Deletion is a point of friction, which demotivates new editors without sufficiently informing them about Wikipedia's values and standards. Highly evolved to meet the needs of the encyclopedia, deletion is an intricate administrative process, which, though common, is poorly understood by readers and novice contributors. Many readers are shocked to learn that Wikipedia deletes articles, and some new editors first learn about Wikipedia's quality standards and the deletion process when an article they wrote is removed. Retaining these editors is more challenging, particularly for the large percentage (~33%) of novice editors who begin editing by creating new articles[1].

Educating and socializing newcomers is vital, both to ensure quality decisions [23] and to equip readers to become contributors [1]. Wikipedia has extensive documentation about policies and content standards, yet the very volume of this advice, as well as its deliberate separation from the content space makes it inaccessible. For recruiting, instructing, and retaining new editors, this is ineffective: failing to inform readers about what is valued in encyclopedic content does not promote good first contributions. This risks alienating potential and new contributors while fatiguing existing ones (who must weed out not only deliberate vandalism but also misguided but well-intentioned contributions), and wastes novices' time in writing and reviewers' time in deleting inappropriate content. To inform readers two qualities are key: succinctness and proximity to the encyclopedic content of the article space.

With this end in mind, we investigate the deletion process at a more granular level of detail than existing research (e.g. [32, 16, 25, 23]), describing the overall deletion process, examining the decision factors and verifying the minimal standards for keeping borderline content and investigating newcomers' experiences in deletion debates. We show that with four factors, over 70% of deletion debates can be decided. This, we believe, offers opportunities for acculturating new editors to reduce the friction of deletion.To open up the encyclopedia, recent efforts have included content rat-

---

[1]"only 0.6 percent of those whose articles are met with deletion stayed editing, compared to 4.4 percent of the users whose articles remained", http://enwp.org/Wikipedia:Wikipedia_Signpost/2011-04-04/Editor_retention

ing mechanisms; we contend that by exposing these four factors as simple decision criteria, the work of deletion can likewise be made more democratic. Since readers can become future contributors, making the desirable criteria clear and visible should help improve the articles contributed while opening up the decision-making process. We next describe the deletion process and present our three research questions.

1.1 The Deletion Process

Wikipedia uses a staged process for deletion, with four main paths, that we have identified and outlined in Figure 1. Relatively less community input is required for clear decisions, and additional participation is encouraged where policy could be interpreted in multiple ways[2]. Clearly inappropriate content (e.g. for copyright violations or attack pages) can be deleted immediately by an administrator in the "speedy deletion" process. Other material considered uncontroversially non-encyclopedic can be nominated under "proposed deletion" (or PROD), which has a 7-day waiting period, during which anyone can contest the PROD by merely editing the page. Finally, controversial cases (~12% [25]) are sent to special community discussions called "Articles for Deletion" (AfD)[3], where users debate the merits of each case, presenting reasons for deleting or keeping an article or otherwise managing its content. Although AfD discussions are open to anyone – even IP users without a username – to read and to comment on[4], few newcomers participate [16], and as the diagram shows, the procedures are complex. After 7 days, AfD discussions are reviewed by an uninvolved user or administrator who attempts to determine the consensus. Rather than closing the discussion, they may relist a debate for further discussion. Otherwise, the 'closer' rules on whether an article should be kept, deleted, merged, redirected, or transwiki'd, based on Wikipedia's rules and policies and the points made in the discussion. Readers are most likely to encounter deletion through markers (Figure 2) rather than policy.
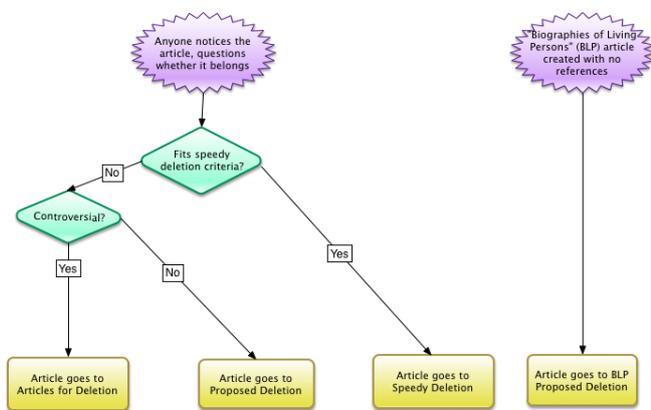

Figure 1: The four types of article deletion in Wikipedia.

Focusing on AfD, the most demanding of the deletion processes in terms of user time and attention, this paper investigates three

---

[2]Based on Wikipedia policy documentation, we created a detailed workflow diagram of the process, available at http://meta.wikimedia.org/wiki/File:Deletion_process_on_English_Wikipedia_(flowchart).jpg. It was verified by experienced users and administrators as of July 2011.

[3]http://enwp.org/WP:AFD

[4]Nominating an article to AfD does require a username.

related research questions. Based on a corpus of deletion debates[5], we investigate:

1. What factors contribute to the decision about whether to delete a given article?

2. When multiple factors are given, what is the relative importance of those factors?

3. What are the outcomes of deletion discussions, both for articles and for the community?

In this work, we discovered that 70% of deletion debates can be completely decided based on four factors: Notability, Sources, Maintenance, and Bias. Further, the complexity of debates is often due to the presence of 'discussions within discussions', many of which are tangent to the overall purpose, and may, for instance propose policy changes, make personal attacks, or provide instructional and procedural advice. Some newcomers who participate in deletion discussions may quickly evolve their notions of Wikipedia's standards, but for those who fail to grasp or who do not agree with the standards, the process appears frustrating.

In this introduction, we have described the deletion process and presented our three research questions. Next we discuss related work. Then, after describing our methodology, we proceed to discuss each research question in turn, covering decision factors, their relative importance, and the article improvement process. We then discuss the results, highlight additional related research, and conclude with a discussion of our own future research.

2. RELATED WORK

We briefly review related work, which falls into several areas: deletion; discussion, argumentation and controversy; policies and values; and novices vs. experts.

2.1 Deletion

Existing research on deletion has focused on shallow analysis of large datasets, e.g. of redacted content [34], vote sequencing [32], and decision quality [23]. By contrast, our study is a content analysis of a focused sample.

Like [25], we investigate the reasons for article deletion, but at a different scope and level of granularity. We do not handle the routine deletions which their keyword-log based study classifies. Rather, the discussions that require significant community input and attention are our focus; while they merely classify these as 'PROD/AFD/VFD', we hand-code factors for deletion (RQ1) within individual Articles for Deletion cases. One of our key deletion factors (RQ1 & RQ2), notability, receives deep treatment in their study, as they investigate the extent to which notability standards changed from December 2004 to March 2008; to this end, they approximate notability with popularity, measuring both page views and search engine results.

Further contributions from Lam et al. [23] center around decision quality, a topic we do not address in this paper. Lam et al. use reversed decisions as indicators of poor decision quality on a corpus of AfD's from January 2005 through April 2009. Further, they analyze the 4.67% of delete decisions (68% of discussions are deletes) and 3.52% of keep decisions (25% of discussions are keeps) that were reversed. They also find that the best decisions are made by larger groups (with diminishing returns; and in their sample the median number of participants is five); that recruitment, while biased, did not have a significant effect on the decision quality; and that

---

[5]See http://jodischneider.com/pubs/data/wikipedia

Figure 2: Markers of deletion include (a) Notices of possible deletion, such as this AfD discussion in progress, or (b) Notices of previously deleted pages.

better decisions arise when administrators close discussions contrary to their own personal keep or delete bias.

In this study, by identifying newcomers and tracing their participation in discussions, we provide evidence of some of the challenges the community faces in socializing newcomers, as part of the community impact of deletion discussions (RQ3). Newcomer participation in deletion discussions has received some attention in previous research. Users with fewer than 500 edits are more likely to argue for keeping the article, and AfD decisions made with the participation of newcomers were more likely to be overturned [23]. Geiger and Ford studied all deletions from June 2007 to July 2011, finding that new users' participation was rare: only 26% of discussions had a newcomer, with only 8% having more than one newcomer [16].

2.2 Discussions, argumentation, and controversy

Our work contributes to a understanding of how groups make decisions. Decision-making discussions in a variety of media have received significant attention, including studies of decision-making in open content discussions. For example, Barcellini has studied how design decisions in the Python community are mediated by boundary-spanning between user- and developer-oriented mailing lists [5] while Ko and Chilana have studied contentiousness in bug reports in several open source software communities [20].

Wikipedia's discussion spaces are considered in a large body of research[6]. Some recent literature identifies the purpose of Talk page discussions (e.g. article criticism, explicit performatives, information content, interpersonal) [12] and, towards finer-grained understanding, annotates authority claims and alignment moves in these discussions [6]. Studies have also shown how discussion spaces are used for consensus-seeking [21] and how argumentative discussions contribute to article improvement [14].

2.3 Policies and values

By investigating the decision rationales Wikipedians articulate, this study contributes to understanding Wikipedia's policies and values, an area that has been extensively studied. Beschastnikh et al. classify policy citations as relating to attribution, consensus, bias, disposition, writing style, genre, inclusion, or legal [7]. Morgan et al. study the rhetorical and value appeals made in a contentious Talk page discussion [26]. Nagar views a policy discussions through the lens of sensemaking, showing how the wiki environment helps stabilize the commitment to and interpretation of the policy [27]. Wikipedia's content policies [9] and grading of content [31], and the community's policy development processes have also received attention (e.g. [21, 13]).

2.4 Novices vs. experts

There is increasing interest in the demographics of Wikipedia, and especially, concern for the editing experience of new Wikipedians [33]. In related work, we are investigating the role of experience and expertise in justifying the deletion decisions made in our corpus [29]. Previous work has found differences in novice needs [4], has shown how editor experience impacts quality [18], and has suggested balancing procedure- and content-oriented members [2].

3. METHODOLOGY

Next we discuss the methodology, detailing our corpus; our iterative coding procedure; the codes ultimately used; our use of article history logs; and our use of user contribution history logs.

We used all 72 debates begun or relisted on January 29, 2011[7]; this is a typical day, with an average number of debates[8]. Previous research categorized all deletions based on keyword logs of all deletions, using rough categories: inappropriate content, no content/context, notability/significance, PROD/AfD, wiki maintenance, other, and unknown [25]. By contrast, we analyze AfD discussions and outcomes based on hand analysis of our sample. The scope of the sample enabled extensive manual examination; basic information about the debate input and outcomes are shown in Figure 3.

For each debate, we analyzed the factors mentioned and we tracked the debate participants. We first began by applying an open coding procedure. Next we compared this to Stvilia's information quality assessment model, developed in the context of article promotion and demotion on Wikipedia. This model maps between three sets of evaluation criteria for encyclopedia articles: Stvilia's own model [31], Wikipedia's model, and Crawford's model for encyclopedia evaluation in traditional media [11]. We devised ten factor codes, then a single coder used this classification to recode our sample, counting each factor no more than once per debate. In descending order of prevalence, these factor codes were as follows: Sufficiently important, content is verifiable, maintenance issues, genre suitable, size of article, topic covered elsewhere, meets minimum requirements, clear topic, and aids comprehensiveness.

Next, these ten factors were used to recode the sample at the individual comment level (multiple factors were allowed per comment); since interannotator agreement between the three annotators was weak, we refined and truncated the factor manual in discussions with annotators. In particular, we used the four most common categories, which covered 91% of comments. In doing so, the earlier 'maintenance issues' was expanded to include issues of topic and information loss. Further, non-arguments (such as 'agree with [USER]','per nominator') along with non-sequiturs and relisting notices were categorized as not expressing an argument. And finally, a catchall 'Other' category was created. Two annotators then

---

[6]http://www.citeulike.org/group/13905

[7]http://enwp.org/WP:Articles_for_deletion/
Log/2011_January_29
[8]based on 2007-present

reclassified the argumentative comments as shown in Figure 4 into one or more of five categories–Notability, Sources, Maintenance, Bias, Other, resulting in good interannotator agreement as shown in Figure 5.

Our classification was based on what Wikipedians wrote. Votes left with no rationale were excluded as having no applicable factors. Our five categories are distinct but not independent. We distinguished interlinked factors, such Sources and Notability. Thus, even though Sources are frequently used as supporting evidence for Notability, we separated the discussion as much as possible. So for instance 'no reliable sources' was coded as Sources; 'not notable' was coded as Notability; 'no reliable sources to indicate notability' was coded as both.

For this study we also collected edit history logs for the kept and redirected articles; we counted how many revisions were made to the article since the AfD notice was added, both overall (through early September 2011) and during the AfD debate (i.e. until the article was edited to remove the AfD notice and indicate the AfD closure). We also used AfD participants' contribution history, identifying participants with fewer than 105 contributions[9] before the beginning of the first AfD debate in our sample (January 14) as novice contributors.

Outcomes	Number of AfD
delete	37
keep	21
no consensus	4
redirect	6
speedy delete	4

Input sources	
contested PROD	8
relisted	16

Figure 3: Outcomes and inputs from our corpus of 72 debates.

	Bias	Maintenance	Notability	Sources	Other
observed agreement	0.95	0.87	0.88	0.92	0.83
Cohen's kappa	0.66	0.64	0.75	0.82	0.34

Figure 5: Observed agreement and Cohen's kappa.

4. DECISION FACTORS

Each discussion mentioned from one to all five of the factors, Notability, Sources, Maintenance, Bias, and Other. The overall frequency of factors based on debate outcomes is shown in Figure 6.

The same factor can be used to argue both for keeping and for deleting an article, as shown in Figure 4. Often a factor is mentioned as a concession to agree with the evidence but disagree with conclusions. In this example, despite maintenance and bias concerns, a notability guideline (*WP:POLITICIAN*) is used to argue for keeping an article about a politician: *Article should definitely be rewritten (sounds promotional) but I think he passes WP:POLITICIAN*. As this demonstrates, factors can offset one another. As in Wikipedia's general discussion spaces, policies and guidelines are of particular importance in articulating consensus views [21]; this causes difficulties for newcomers, in articulating effective arguments, as we further discuss in Section 5. Even

[9]This was a natural dividing point; we experimented with distinguishing brand-new and relatively inexperienced users but this didn't make significant differences.

though numerous policies and guidelines are cited (and particularly the subject-specific notability guidelines for applying the notability policy to professions, organizations, songs, etc.), the arguments made in 69.5% of discussions and 91% of comments are well-represented by just four factors: Notability, Sources, Maintenance, Bias.

4.1 Number of factors

The number of factors might be expected to correlate with the level of controversy of a debate and its length. Indeed, unanimous decisions[10] were shorter and indicated fewer factors. Fully one-third of the sample–24 articles–was deleted unanimously, with no opposition to deletion. These debates had were two to ten comments (median 6.5, mean 6.17) and mentioned one to four factors (median 2, mean 2.33). By contrast, non-unanimous discussions (contested deletions, kept, or redirected) were typically far more voluminous, with five to thirty-three comments (median 11, mean 12.35), and most (93.75%) non-unanimous discussions mentioned multiple factors.

Yet contested deletion debates mentioned more factors than keep debates. This is surprising since, in the main, keep decisions and contested deletion decisions are of similar length, with at least six comments (means 12.05 keep, 11.44 contested deletion; medians 10 kept, 10.5 contested deletion). In the extreme, the longest keep discussions in our sample received thirty-three comments, compared to twenty-one comments for a contested deletion. In both cases, controversy is a given, but the nature of the discussions varies somewhat.

Discussions for contested deletions[11] were fairly similar to keep discussions in their broad outlines but on average, contested deletions mentioned more factors (mean 3.39 for non-unanimous deletions, compared to 2.71 for keep), and this difference is stable when the lowest and highest-factor debates in each category are truncated.

4.2 Relative Importance of Decision Factors

The relative importance of decision factors is shown in part by their prevalence, shown in Figure 7, and is also indicated by how salient they were in debates. Decision factors for keeping, deleting, and redirecting articles differed, as shown in Figure 7(a). Bias was the least mentioned, and least salient: Bias alone never closed a debate, while each of the other four factors was the sole factor deciding one debate[12]. 'Other' issues, while significant, were a minor point of discussion, except in Redirects, where they made up 28% of the discussion. In Speedy Deletes, Notability, Sources, and Bias were key points of discussion, while aintenance was not discussed, and 'Other' issues were particularly insignificant.

Notability, the importance of content, is the most frequently raised factor in discussions, and often the most decisive factor. A compelling argument for notability can close a debate decisively, when there are not other issues at stake: a single argument, publicly accepted by multiple participants can decide the case (e.g. for Melqui Torres). A very typical argument is that sources are needed to show the notability of a topic, or that the subject does not meet a specific notability guideline. Yet Notability does not always carry the debate: *we shouldn't mechanically apply notability guidelines*

[10]Note that kept articles are never unanimous: kept articles have disagreement at least with the nomination.

[11]Including one speedy delete that was contested by its author and subject.

[12]Notability and Sources were each the sole factor in a No Consensus close; these cases, which received insufficient comments; they are discussed further below in Section 5.1.1.

factors	Example (used to justify `keep')	Example (used to justify `delete')
Notability	Anyone covered by another encyclopedic reference is considered notable enough for inclusion in Wikipedia.	There is simply no coverage in reliable sources to establish notability.
Sources	Basic information about this album at a minimum is certainly verifiable, it's a major label release, and a highly notable band.	There are no independent secondary sources (books, magazine articles, documentaries, etc.) about her.
Maintenance	this article is savable but at its current state, needs a lot of improvement.	Too soon for a page likely to be littered with rumour and speculation.
Bias	It is by no means spam (it does not promote the products).	The article seems to have been created by her or her agent as a promotional device.
Other	I'm advocating a blanket "hangon" for all articles on newly-drafted players	it appears to be original research by synthesis

Figure 4: Decision factors (in descending order of prevalence) can be used to argue for either keeping or deleting an article.

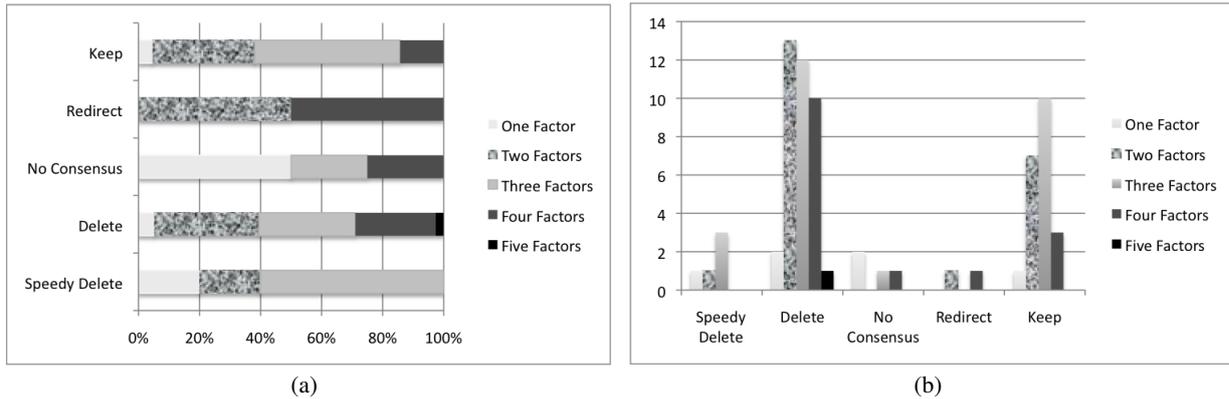(a)   (b)

Figure 6: The number of factors, from Notability, Sources, Maintenance, Bias, and Other, found in AfD decisions (speedy delete, delete, no consensus, redirect, keep), by (a) percentage (b) and number of debates.

in this instance, where it would "[punch a] hole in their otherwise comprehensive discography." Even when a topic's notability is not disputed, it may factor into the discussion, as this closing summary emphasizes: *this is the rare case where notability is not the main argument in favor of deletion. It has been demonstrated that the subject is already covered in numerous other articles and that those articles do a much better, more thorough job of covering the topic.* Structural issues beyond the article level, such as comprehensiveness, or other articles that could cover the content, are the most effective arguments that can trump Notability.

Sources and Notability are distinct, yet closely related factors. Frequently Sources are used to support claims about Notability; a typical argument is that, when no sources are found, notability cannot be established. However, the Sources factor can be discussed without explicitly mentioning, or even necessarily implying, a particular Notability judgement, for instance to analyze the reliability of a given source or to indicate that sources had been added since the nomination.

Maintenance concerns sometimes contribute to deletion discussions. Articles for deletion often serves as a place for pruning or sorting content that occurs in or might fit in multiple places. Even once content is determined to be encyclopedic, the need for a particular article is not assumed. Then, only the amount of content decisively saves the article: "List of legislation sponsored by Ron Paul" was kept in part due to *too much content to merge in to another article*; precedent was also on the side of keeping: *this article is a breakout of the WP:SUMMARY in Ron Paul#Legislation, as agreed since 2007 to manage this degree of notable content.* Notable content, however, could be deleted altogether due to mainte-

nance concerns, especially about the topic: *not worth rescuing such a vague concept.* It is rare, but possible, for Maintenance to be the only issue in a debate: in our corpus, the Maintenance delete was an unusual case of long-standing duplicate content. Some maintenance issues, such as content sorting, cannot always be effectively dealt with in isolation at the single article level.
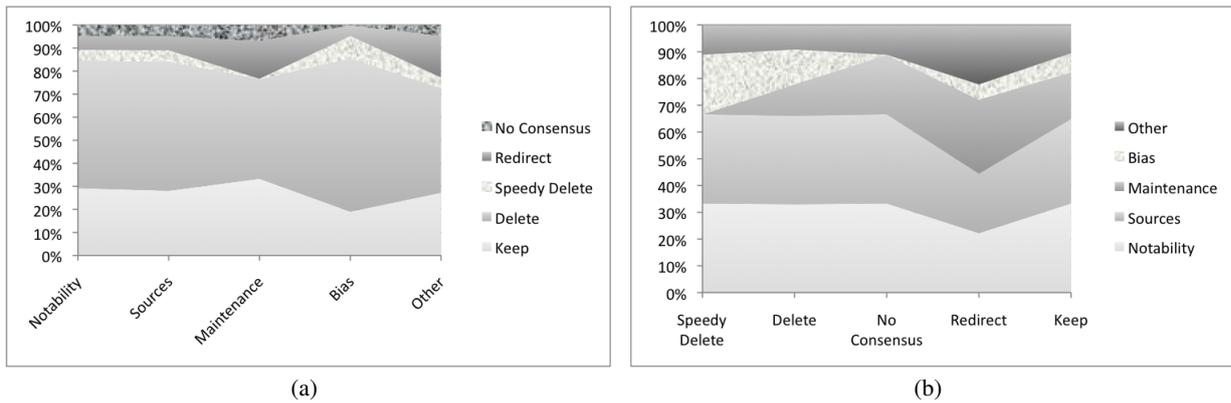
Bias was not a deciding factor; it made decisions about non-notable content swifter and less contentious, sometimes leading to Speedy Deletes that forestalled the need for further discussion. Yet notable content not covered elsewhere was kept; Figure 8 shows an example in which Notability trumps Bias. The lack of independent, reliable sources, however, often accompanied bias: if sources could not be found, the article could not be kept (and could also not be deemed important, since that relies on external validation).

The most frequent 'Other' concerns were genre and copyright; copyright issues can close discussion immediately, since inappropriate content can be deleted under CSD without discussion. Concerns about the suitability of the genre were decisive if there was no potential to write a more appropriate article on the topic. Yet genre issues could sometimes be countered by the ability to improve the article; for this reason, *Wikipedia is not a book report* does not carry the debate extracted in Figure 8.

Decisions of no consensus rested on the Notability, Maintenance, and Sources available for content. We discuss these in further detail next, when we examine the outcomes for articles and for participants.

5. OUTCOMES
The deletion process clearly impacts articles – determining

(a)

(b)

Figure 7: **(a) Prevalence of factors based on debate decision: no consensus, redirect, speedy delete, delete, and keep; and (b) Percentage of debate decisions associated with factors: Notability, Sources, Maintenance, Bias, and Other.**

Sources, Notability	Read likes an **WP:OR** book report, only two citations, **no content about why book is notable**
Bias	Agreed... Also somewhat **biased in tone**. Merge and Redirect
Other [Size], Maintenance	Keep. This article has been in existence since 2004. It is **not some little stub article**, either. **If you don't like the way the article is written, then fix it**. ...
-	Also, a merge and redirect does not mean the content will be deleted, just included with the Randy Barnett article. ...
Other [Genre]	Merge and redirect **Wikipedia is not a book report.** If this book received media coverage and commentary, write an article about that.Shii (tock) 02:09, 31 January 2011 (UTC)
Notability, Sources	Keep -- **this book been much discussed in U.S. constitutional scholarship in recent years, and has won awards**, per existing source. The article clearly **needs more sources for it's claims**, but article deletion is inappropriate...
Sources, Bias, Notability	Merge and redirect : The entire 'theory' section is **unsourced original research** and/or **point of view**, and should be deleted immediately. **There are no sources on the page that assert notability for the book even if it is an award winner.** Verifiability, not truth, is the core policy of Wikipedia. Kudpung (talk) 01:55, 3 February 2011 (UTC)
Notability	Keep: Barnett's a bit kooky, but **i see indications of notability** that shouldn't be ignored. **Book review in a major law review journal: [7]; book review in American Prospect: [8]; review in washington times [9]; mention in passing by Jeffrey Rosen (leading legal writer) at NYT[10] called it a "provocative book".** From most of these sources you can predict Barnett's political leaning, but these aren't petty sources.--Milowent • talkblp-r 05:06, 4 February 2011 (UTC)

Figure 8: Extracts from an Articles for Deletion debate, tagged with codes from our final codebook. Statements in bold are related to the codes on the left; each comment is in a separate row. Ultimately the article was kept.

whether an article or its content are deleted or retained, and potentially drawing attention and editing to the article. Participants, particularly novice participants, are also impacted by their participant in the deletion process, and by the deletion of content they read or contribute.

5.1 Decisions of No Consensus

In our no consensus sample, there are two types of no consensus decisions: first, those which failed to gain sufficient discussion and second, those in which the article creator (and sometimes others) made changes to the article during the discussion period. Our corpus contains four No Consensus decisions, two of each type, which we next discuss.

5.1.1 Discussions fail without comments

One enduring issue is bringing sufficient participation to deletion discussions. Overall participation in AfDs is limited; from January 2005 through April 2009 the median number of participants was five [23], and from June 2007 to July 2011, 83.62% of all AfDs had between 4 and 12 participants. Discussions that failed to reach consensus are particularly relevant, since they show some of the problems with the process: 'No Consensus' outcomes are more likely to be result in articles being renominated for deletion later.

Lack of discussion also implied lack of editing to the article. Discussions do not always attract enough participants. We now discuss the two articles in our corpus in this situation.

RobApps.

RobApps failed to attract any comments beyond the deletion nomination, despite being open for comments for 31 days (more than 4 times the norm) and relisted at AfD twice. Yet in a subsequent AfD (when it was relisted for deletion March 26), the unanimous consensus was that it was spam and a borderline candidate for speedy deletion.

Norazia.

Likewise, the unsourced biography Norazia spent over 3 weeks at AfD, failing to attract attention until the third time it was listed, when both English and foreign language sources were proposed. This was a minor success for the AfD process and for the Biographies of Living People sourcing project [19]: one editor attracted by the AfD added sources. Yet the article failed to attract further attention; of its 25 edits since creation in July 2006, five took place

in conjunction with the AfD process, since then there have been no further edits: the most recent revision, by the debate closer, indicates the no-consensus close of this debate. Worsening the situation, from a quality perspective, many of the intervening edits were small technical changes, and a bot [15] is the top contributor.

5.1.2 Interactions with Article Creators

Interactions with the article creator are characteristic of the other two no consensus decisions in this sample, William Vickers (fiddler) and St. Andrew's Episcopal School (Amarillo, Texas). The AfD process had significant effects on these articles, which focused their editing and provided feedback to the main contributors–feedback which was lacking in the article Talk page, the normal discussion venue. The only comments on either Talk page are spillover from the AfD discussion.

William Vickers (fiddler).

Nominated for deletion 1 hour and 20 minutes after its creation, William Vickers (fiddler) has had few edits outside its main author; others made five of its 43 edits during the AfD process and mainly as a part of that process. Yet the AfD process shaped this page. The author's contributions are certainly more voluminous due to the AfD. This was the first of eleven articles created by this author: only one has more than nine contributions from him (it has 26), and many have as few as three of his contributions.

Suggestions made in the AfD were implemented in the page. First, they led the author to rename the page focusing on a more appropriate topic: the manuscript rather than the man who *Little is known of*. Second, in response to a call for further sources, the primary author added a discography. Although similar discussions could have happened on the article discussion page (this article still has none, barring a link to the now-closed AfD), immediate feedback (which came not long after the article was created, in the first 3 hours after the deletion nomination) was probably helpful to the article development. The length of the debate period may also have been a factor: the no consensus decision is in part due to lack of comments when the debate was relisted, twice, for further discussion.

St. Andrew's Episcopal School (Amarillo, Texas).

While mentoring a new contributor was also a feature of the AfD for St. Andrew's Episcopal School (Amarillo, Texas), there was far more negative emotion. Its importance, or notability, was the main issue of contention: Except in unusual circumstances, elementary schools are redirected to the corresponding high school. The primary question, then, was whether this school (which, as an independent school with no district or high school, lacked an obvious redirection target) was sufficiently notable on its own.

The contributor's behavior, not just the article, came under discussion: s/he had marked other articles for possible deletion in the PROD process, garnering a cynical response: *Maybe it's in bad taste but if my school does not meet WP standards then why should others??* This was followed up by a message indicating discouragement: *To be honest it's been a real turn off adding articles to WP and I don't think I will add articles again. So smile and enjoy.* Only the persistence of an advocate for the novice, who co-edited and argued strongly for the article, ameliorated the situation.

This was the third article the user created, all within a single week. Again, AfD helped increase contributions: as opposed to three or four contributions from this user, s/he has made 18 edits to the article, which has received 53 edits overall (38 during the AfD process itself).

Previous research has found that creators rarely (17.59%) discuss

the deletion of their articles [16]; encouraging positive interactions with creators should be a design goal of future development. In our corpus, negative interactions were mainly due to conflicts around Wikipedia's consensus values; article creators who do not understand these values express frustration with the process. In extreme cases, creators are banned from these negative interactions (this happened once in our corpus, with a novice editor whose autobiography had inherent sourcing problems). We next discuss conflicts around consensus values.

5.2 Conflicts around Consensus Values

5.2.1 Emsworth Cricket Club

Emsworth Cricket Club, deleted due to lack of notability, garnered 18 comments from 15 AfD participants, including 3 IP users, and 2 newly-registered users. Discussion was completely split: all the novices argued for keeping, while none of the more experienced users did so.

The discussion focused on notability, and keep arguments showed novices' confusion about Wikipedia's scope and purpose: *Why just because it is a small team and not major does it not deserve it's own page on here?*. Novices' confusion seems not merely about the importance or notability criteria of Wikipedia, but rather seems to be about whether there are minimal standards, when there are other articles about *pointless people* such as celebrities and *a breed of dog*. Their comments are poignant and extensive, with high emotion and many analogies.

This debate was a failure insofar as its outcome was clear, yet there was little constructive engagement between novices and experienced users. The new usernames seem to have been registered just for the purpose of participating in the debate, as they have not been used for editing.

5.2.2 Prometheus (film project)

This redirect article was debated mainly due to maintenance (its small size, existing coverage elsewhere) and sources, as an article *likely to be littered with rumour and speculation* because it was *too soon*. Yet policy was the main issue underlying the redirect decision of this article: *we do not treat "film projects" as films until principle filming actually begins*.

This debate was a failure insofar as it involved considerable effort–28 comments from 13 users. However, it may have increased some participants' policy knowledge, and presumably increased the maintainability of the article (which has been recreated, now that filming is underway). Benefit to the article is harder to measure; half of the 11 edits made during the AfD were made by debate participants. The debate attracted one new editor who made two edits during the AfD and the article's second most active editor made three edits. Yet this is little more than the weekly average number of edits (283 total) over the article's lifetime.

While the outcome was unsurprising to those versed in policy, the policy itself came under attack in the discussion: *Frankly, the basis of my disagreement with you here is that I don't agree with the guideline*. Participants tried to argue for the importance and maintainability of the article, as well as (inappropriately) for its popularity *obviously of interest to the public in general*. This points to the need for skill in debates: along with understanding what factors are relevant and which (e.g. popularity) are not considered relevant, participants need both knowledge and acceptance of policy. This debate was not unique in generating discussion about policies and we next discuss attempts at policy development that happened within AfD.

5.3 Policy Development

Policy is often debated or proposed in Articles for Deletion. This adds to the length and complexity of a debate, but there is little indication that it impacts either policy development or the debate outcome.

5.3.1   Policy development is not permitted

Policy development may start in article discussion pages and continue in Articles for Deletion. One such example is the case of Bryan Meredith (soccer), where an attempt at policy development was met favorably by discussion participants: the proposal *I'm advocating a blanket "hangon" for all articles on newly-drafted players* was met with replies agreeing to *wait for a while*. Yet AfD is not recognized as an appropriate venue for such policy development.

The debate was closed as Delete, and the extensive discussion of a policy proposal led to this remark from the closer: *If you wish to install such a policy, do so on a discussion page; do not make up rationales contrary to policy and guidelines on individual AfDs.* In otherwords: take it elsewhere. This again is a learning opportunity, as well as an enforcement of procedures over content. Policy development and policy application are not tightly integrated, and there is a tension between contextualizing discussions and separating discussions into policy spaces.

5.3.2   Increases the complexity of the debate

The length and complexity of a debate depend on how persuasive the first respondents find this argument; a single argument, publicly accepted by multiple participants can decide the case, but the order of replies matters to the character of the discussion and impacts what issues get raised [32].

Two debates about baseball players, Melqui Torres and Heath Totten, show this compellingly. Nearly identical proposals were made about why these players were 'non-notable', and in both cases the first response is a keep argument providing evidence about how the player meets the baseball notability guidelines. In Torres' case, two replies quickly affirm that first argument, but provide no further information. Yet the Totten discussion is nearly twice as long, and while multiple pieces of evidence are given about his notability, there are also sidetracks both about how to apply the criteria of the appropriate notability guideline (i.e. Are particular venues sufficiently important? *The Olympics are "major." The Caribbean Series is not.*), as well as concerns expressed about the guideline itself *That rule really needs tweaking then.* In the end, nearly half the discussion concerns changing the policy.

With nearly identical nominations, and nearly identical responses, these debates show that the order of replies, and the presence of related deletion discussions have an impact.

6.   DISCUSSION

Differences in editing behavior are apparent from a user's first edits [28], making it even more vital to make reading the encyclopedia a learning experience about editing it [1]. Currently, readers are most likely to encounter deletion through notices of ongoing deletions such as Figure 2(a) (or analogous notices for PROD and BLP-PROD deletions) and indications of previously-deleted pages (Figure 2(b)). These provide little indication about what is important, though curious readers can follow links into policy and guideline documents and previous discussions. This is a severe problem with large impact: currently 22% of all deletions are speedy deletions of articles that do not indicate their importance [16]. Our work contributes to addressing this problem by indicating that very few content standards need to be clearly communicated to readers in order to bring significant benefit. 69.5% of discussions and 91%

of comments are well-represented by just four factors: Notability, Sources, Maintenance and Bias. The best way to avoid deletion is for readers to understand these criteria.

Our position is that readers become writers, and hence, affordances that inform readers peripherally are essential. These affordances can take various forms. First, existing markers, such as those shown in Figure 2, should be regularly reviewed; small changes in wording and linking strategies can make a significant difference in promoting newcomer understanding [17]. Second, incentivizing simplification and clarity of guidelines could help. What if we wrote policy summaries in simple English, and promoted a policy of the week on the Main page? Third, tools aimed at novice contributers, such as the Article Wizard[13], play an important role. The Article Wizard steps new creators through the process. It reduces the problem of deletion by identifying in advance the reasons an article might be deleted, and presenting policies in bite-size chunks at appropriate moments. In effect, it translates policy into personalized decision-support in context. Fourth, reader-oriented tools should be aligned, to the fullest extent possible, with the language and policy that editors use. Again, the words used are key.

Terminology issues make it difficult to instruct newcomers as to why their articles are being deleted, or why articles submitted with the Article Wizard are not accepted. For instance, a statement that there are 'no reliable sources' must be taken in context. For an experienced Wikipedia, that encapsulates an entire policy, as well as an argumentative dialogue about whether, in this context, a source is appropriate for a particular purpose.

The same vocabulary should be used by the encyclopedia's editors and its readers. We think that the factors we have identified – Notability, Sources, Maintenance, and Bias – address fundamental high-level criteria for content. This is common terminology within Wikipedia; articles for Notability, Sources, and Maintenance, and the Bias disambiguation page currently point to community areas for these topics within the Wikipedia namespace. Yet these are not the words we are choosing to teach readers.

In particular, the most prominent deletion indicator – notability – is nowhere communicated to readers. This is a problem that could be fixed–thus raising readers' awareness of what is important for writing the encyclopedia. To address notability, readers could be asked to flag pages that are do not discuss important topics, or asked what aspect of a page's topic makes it worthy of inclusion in the encyclopedia.

Currently, there is a significant–and we argue, problematic–divide between the language promoted to readers and to editors. For instance, with the Article Feedback Tool[14], Wikipedia readers are asked to rate pages as trustworthy, objective, complete, and well-written. These relate to, but do not directly correspond with, the criteria by which articles are judged for deletion. Nor do they precisely match the terminology used in Featured article criteria. This is a lost opportunity for informing readers–who are potential contributors–about the community values and standards.

7.   FUTURE WORK

Decision-making based on open contributions can be challenging, with imbalances in power between those providing opinions and those implementing decisions [20]. Voluntary, collaborative communities require robust mechanisms for sharing decision-making criteria, since implementation may be a distributed, and perhaps temporary task. Thus establishing shared mental models

---

[13] http://enwp.org/wiki/Article_Wizard
[14] http://www.mediawiki.org/wiki/Article_feedback

(e.g. for new participants) about standards is key to efficient, collaborative maintenance. Wikipedia, as a well-developed voluntary, collaborative community, is a good place to examine decision making. Models developed on Wikipedia may be relevant for other communities as well: similar issues arise in open source bug reporting and merging content in question/answer communities such as StackOverflow.

Further, understanding current biases and benefits of the process can be helpful for tweaking the process, or providing dedicated support in decision-making. Deliberation and decision-making support tools can be customized to particular collaborative work environments such as Wikipedia. Such support can be used to defuse contentious conversations while also socializing newcomers (e.g. [8, 22]). By crystalizing current processes, such tools increase transparency, but at the risk of losing flexibility for easily changing the process. By focusing first on the enduring decision factors [3], we bring increased understanding to the process.

Currently, decisions about which articles to delete are left to long-standing users by default when not by policy: few newcomers participate [16]. When procedural knowledge about Wikipedia is valued more than content expertise, this limits the effectiveness of decisions and risks deleting content that is worthwhile but not well-defended.

We have mainly focused, in this paper, on the risks and benefits to newcomers. However, deletion offers pain points on all sides. As the Wikipedia Signpost recently commented[15], long discussions "often present a challenge to the editor who steps up to close them; 'no consensus' is a common outcome for convoluted debates, a lack of resolution that opens the possibility of discussion starting all over again as the same issues continue to arise." Based on the decision factors presented in this work, we preparing a tool aimed at debate closers, that summarizes the decision factors in a debate and organizes comments by decision factor [30].

To extend our decision factors visualization to debates beyond our annotated corpus, decision factors could be added manually by debate participants, or automatically determined. We plan to test both techniques; for the latter, we will use our content analysis as an annotated corpus for machine learning, to detect the main issues at stake. Automation results can be evaluated first with closers' decision rationales for archived debates, and then tried as a support for live debates.

There is further potential for support in other aspects of the deletion debates. For instance, archived debates could benefit from audit and summarization tools. Audit tools could help identify decisions worth reviewing. Overturning a decision does not always indicate that a decision was bad. Topics may be unacceptable at one time, and later acceptable. Some of these may be identified by statements that someone is close to meeting the notability condition, or by the particular policies applied (e.g. No Future Films, for Prometheus (film project)). Audit tools for archived deletion decisions could help editors identify material that might now be within scope for the encyclopedia.

Another desideratum for archived debates is to provide readable summaries even when discussions are long and complex. Towards this end, we would like to identify and distinguish the "conversation within a conversation" subthreads that add complexity to long debates, using methods for detecting topic divergence, as well as new methods focusing on user needs and emotions.

Social sensitivity appears to be a key factor in shortening and smoothing discussions, both with novices and with other experienced editors. In future research we are preparing a taxonomy of

the emotional triggers and the associated needs discussants are trying to address (e.g. 'understand why this article was deleted', 'provide further information about a point that was not taken into consideration', 'vent about policy and bureaucratic challenges'). Such a taxonomy, along with the factors analysis given in this paper, could help provide needs-appropriate support. Understanding common editors' needs could help suggest ways to meet participants' needs while defusing emotional debates, and might suggest likely subproblems that could be fruitfully addressed.

Retaining and best using the skills of novice editors depends in part on socializing them to the community norms. Article quality and appropriateness of content are key aspects of this socialization. The Article Wizard points to one possible direction: creating a wizard tool to support novices. Another direction is helping the community identify mentoring opportunities, especially with new creators, in deletion and in other venues. Deletion is a high stakes venue, and harnessing the community's time and attention in AfD could support novices in learning.

Co-editing and prominent discussions of how to improve an article already occur alongside and in deletion debates. Future development should promote the positive interactions with creators–such as those that occurred in the William Vickers (fiddler) and St. Andrew's Episcopal School (Amarillo, Texas). Had these articles been summarily deleted, we doubt that these Wikipedians, as first-time article creators, would have immediately created subsequent articles. Even when a creator's article is deleted, when the overall tenor of the discussion is positive, the result seems to be beneficial for the creator (who gets personalized mentoring on article policy)[16]. Fewer than one in five deletion discussions includes the article creator; we suggest flagging these cases for further attention from skillful mentors.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] J. Antin and C. Cheshire. Readers are not free-riders: reading as a form of participation on Wikipedia. In *CSCW '10*, pages 127–130.

[2] O. Arazy, O. Nov, R. Patterson, and L. Yeo. Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4):71–98, 4 2011.

[3] K. D. Ashley. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals*. MIT Press, Cambridge, MA, USA, 1991.

[4] C. Atzenbeck and D. L. Hicks. Socs: increasing social and group awareness for wikis by example of Wikipedia. In *WikiSym '08*, pages 9:1–9:10.

[5] F. Barcellini, F. Détienne, and J. Burkhardt. User and developer mediation in an open source software community: Boundary spanning through cross participation in online

---

[15]http://enwp.org/wiki/Wikipedia:Wikipedia_Signpost/2012-03-19/Discussion_report

[16]See for instance http://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion/Hugh_Allison.

discussions. *International Journal of Human-Computer Studies*, 66(7):558–570, July 2008.

[6] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media '11*, pages 48–57.

[7] I. Beschastnikh, T. Kriplean, and D. W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *ICWSM '08*.

[8] M. Billings and L. A. Watts. Understanding dispute resolution online: using text to reflect personal and substantive issues in conflict. In *CHI '10*, pages 1447–1456.

[9] B. Butler, E. Joyce, and J. Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *CHI '08*, pages 1101–1110.

[10] N. Cohen. Define gender gap? look up Wikipedia's contributor list. *New York Times* January 30, 2011. http://www.nytimes.com/2011/01/31/business/media/31link.html

[11] H. Crawford. Encyclopedias. In *Reference and information services: An introduction*, pages 433–459. Libraries Unlimited, 3rd ed. edition, 2001.

[12] O. Ferschke, I. Gurevych, and Y. Chebotar. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *European Chapter of the Association for Computational Linguistics 2012*.

[13] A. Forte, V. Larco, and A. Bruckman. Decentralization in Wikipedia governance. *Journal of Management Information Systems*, 26(1):49–72, 2009.

[14] D. Fréard, A. Denis, F. Détienne, M. Baker, M. Quignard, and F. Barcellini. The role of argumentation in online epistemic communities: the anatomy of a conflict in Wikipedia. In *European Conference on Cognitive Ergonomics '10*, pages 91–98.

[15] R. S. Geiger. *The lives of bots*, pages 78–93. Institute of Network Cultures, Amsterdam, 2011.

[16] R. S. Geiger and H. Ford. Participation in Wikipedia's article deletion processes. In *WikiSym '11*, pages 201–202.

[17] R. S. Geiger, A. Halfaker, M. Pinchuk, and S. Walling. Defense mechanism or socialization tactic? Improving Wikipedia's notifications to rejected contributors. In *ICWSM 2012*.

[18] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl. A jury of your peers. In *WikiSym '09*, pages 1–10.

[19] E. Joyce, B. Butler, and J. Pike. Handling flammable materials: Wikipedia biographies of living persons as contentious objects. *iConference '11*, pages 25–32.

[20] A. J. Ko and P. K. Chilana. Design, discussion, and dissent in open bug reports. In *iConference '11*, pages 106–113.

[21] T. Kriplean, I. Beschastnikh, D. W. McDonald, and S. A. Golder. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *GROUP '07*, pages 167–176.

[22] T. Kriplean, J. T. Morgan, D. Freelon, A. Borning, and L. Bennett. ConsiderIt: Improving structured public deliberation. In *CHI EA '11*, pages 1831–1836.

[23] S. K. Lam, J. Karim, and J. Riedl. The effects of group composition on decision quality in a social production community. In *GROUP '10*, pages 55–64.

[24] S. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant,

L. Terveen, and J. Riedl. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In *WikiSym '11*, pages 1–10.

[25] S. T. K. Lam and J. Riedl. Is Wikipedia growing a longer tail? In *GROUP '09*, pages 105–114.

[26] J. T. Morgan, R. M. Mason, and K. Nahon. Lifting the veil: the expression of values in online communities. In *iConference '11*, pages 8–15.

[27] Y. Nagar. What do you think?: the structuring of an online community as a collective-sensemaking process. In *CSCW '12*, pages 393–402 .

[28] K. Panciera, A. Halfaker, and L. Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *GROUP '09*, pages 51–60.

[29] J. Schneider, A. Passant, and S. Decker. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online taskgroups. *Submitted to CSCW '13*.

[30] J. Schneider and K. Samp. Alternative interfaces for deletion discussions in Wikipedia: Some proposals using decision factors. In *WikiSym 2012 [demo track]*.

[31] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.

[32] D. Taraborelli and G. L. Ciampaglia. Beyond notability. Collective deliberation on content inclusion in Wikipedia. In *Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, 2010.

[33] P. Vora, N. Komura, and S. U. Team. The n00b Wikipedia editing experience. In *WikiSym '10*, pages 36:1–36:3.

[34] A. G. West and I. Lee. What Wikipedia deletes: Characterizing dangerous collaborative content. In *WikiSym '11*, pages 25–28.