Page Protection: Another Missing Dimension of Wikipedia Research

Benjamin Mako Hill
University of Washington
Department of Communication
makohill@uw.edu

Aaron Shaw
Northwestern University
Department of Communication Studies
aaronshaw@northwestern.edu

ABSTRACT

Page protection is a feature of wiki software that allows administrators to restrict contributions to particular pages. For example, pages are frequently protected so that they can only be edited by administrators. Page protection affects tens of thousands of pages in English Wikipedia and renders many of Wikipedia's most visible pages uneditable by the vast majority of visitors. That said, page protection has attracted very little attention and is rarely taken into account by researchers. This note describes page protection and illustrates why it plays an important role in shaping user behavior on wikis. We also present a new longitudinal dataset of page protection events for English Wikipedia, the software used to produce it, and results from tests that support both the validity of the dataset and the impact of page protection on patterns of editing.

Categories and Subject Descriptors

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

General Terms

Wikipedia, Peer Production, Page Protection, Wikis

1. INTRODUCTION

Page protection is a feature of wiki software that allows administrators to restrict contributions to particular pages. For example, a page can be protected so that only administrators can edit it. Protected pages can be distinguished from normal pages by the replacement of the "Edit" button with a "View Source" button and, in many cases, by a small lock icon that appears near the top-right corner of the page (see Figure 1). Protection might involve "full protection" where a page can only be edited by administrators (i.e., "sysops"), or "semi-protection" where a page can only be edited by accounts with a history of good edits (i.e., "autoconfirmed" users). Additionally, protection can prevent specific types of contributions such as editing, moving, creating, or uploading.

OpenSym '15, August 19-21, 2015, San Francisco, CA, USA
ACM 978-1-4503-3666-6/15/08.
http://dx.doi.org/10.1145/2788993.2789846

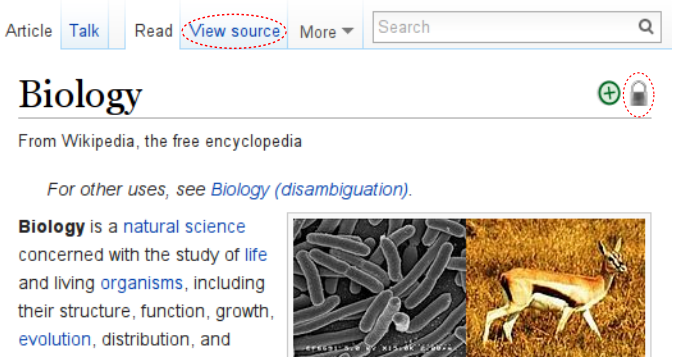

Figure 1: Example of the English Wikipedia article on Biology. Circled in red are the "View Source" button (instead of "Edit") and the small lock icon which each signal that the page is protected.

Page protection profoundly shapes activity on Wikipedia. For example, page protection is an important tool used to manage access and participation when vandalism or interpersonal conflict threatens to undermine content quality. In this way, page protection is a key aspect of the encyclopedia's "hidden order" [11] and contributes to wikis' usability and user experience. While it affects only a small portion of pages in English Wikipedia, many of the most highly viewed pages are protected. For example, the "Main Page" in English Wikipedia has been protected since February 2006. Articles are protected when they are featured on or promoted from the Main Page. Millions of viewers cannot edit protected pages because they never see an edit button.

Page protection has attracted scholarly attention indirectly. It is often invoked in the context of vandalism and anti-vandalism work, as well as related modes of behind-the-scenes, organizational "wikiwork" [5, 7]. Nevertheless, we are not familiar with any quantitative research on Wikipedia that takes page protection into account.

This note makes several contributions. First, we introduce a longitudinal dataset of nearly 300,000 page protection "spells" (time periods during which specific articles were protected) in English Wikipedia between September 2008 and January 2015. We also describe the software used to create the dataset. Second, we use this dataset to characterize page protection activity in English Wikipedia. Finally, we pro-

vide results from validation tests of this dataset and show that accounting for page protection reveals dynamics that researchers should address. We conclude with guidance for how future research might incorporate page protection.

2. DESCRIBING PAGE PROTECTION

Wikipedia administrators protect pages for many reasons. Pages about controversial topics are protected to prevent "edit wars" between contributors with divergent views. Prior work has shown that viewership is positively associated with vandalism and low quality contributions by new users [3]. As a result, popular articles are often protected. For example, consistently high traffic pages like *Biology* may be protected for extended periods. Articles related to current events may be protected while they are in the news (e.g., the article on the February 2014 sports event *Super Bowl XLVIII* was protected during a seven week window in early 2014). Articles "featured" on the front page of Wikipedia are protected as a matter of course. Additionally, protection is used to ensure the stability of pages the community has determined should provide consistent content, like policy pages. Protection is also used by the Wikipedia community and the Wikimedia Foundation (WMF) to enforce policies and the law.

Protection is highly configurable. On English Wikipedia, protection comes in at least ten varieties, including "full protection" and "semi-protection."[1] Protection status can be configured to expire at a predefined time or to last indefinitely. MediaWiki, the software that runs Wikipedia and many other wikis, enforces page protection by preventing actions prohibited by a page's protection status. Protection status can be modified and is subject to discussion and controversy. When the protection status of an article changes, the change is normally recorded by MediaWiki in a log.[2]

Page protection is an important example of how less visible aspects of wikis and social computing platforms shape users' participation and experience. Previous research has analyzed similarly unobtrusive and "hidden" elements of Wikipedia, including aspects of its dispute resolution procedures and coordination work [11, 8]. Studies of peripheral forms of participation have shown that the work of producing the encyclopedia consists of much more than editing the content of articles [1, 2]. Specific types of "wikiwork" have also attracted scholarly attention [7], including recent analysis of other dimensions of the encyclopedia hitherto ignored in most Wikipedia research [4]. None of these studies have examined or accounted for page protection. Page protection is both another form of hidden wikiwork as well as a critical part of the encyclopedia's socio-technical infrastructure.

3. PAGE PROTECTION DATASET

Adopting a strategy similar to one used in our previous work on redirects [4], we have created and published a dataset of *page protection spells* that represent the periods when pages were protected. Table 1 contains four example spells from the article *Biology*. Each spell includes two types of metadata: (1) *type* describing the actions restricted ("edit," "move," or "create"); and (2) *level* describing the class of users to whom that action is restricted ("sysop" means administrators; "autoconfirmed" means established editors; and "templateeditor" indicates a third class of users).[3]

Building a dataset of page protections is challenging because protection information is represented differently and incompletely across different sources of Wikipedia data. One source of page protection data is a "log" of events available through the MediaWiki API[4] and published as an XML database dump.[5] The log records page protection actions alongside other events including page moves, deletions, and editor blocks. In addition to log data, the WMF publishes SQL "page info" database dumps that provide a snapshot of page protection status at the time that the logs and dumps are created. We use both the complete XML log and SQL data to create and validate our dataset.

Unfortunately, the standard for recording page protection data in the log has shifted over time. As a result, log records for page protection events have been recorded incompletely and inconsistently. The most complete and reliable data on page protection events began in late September 2008, when MediaWiki introduced the format for describing page protection log events used today. The spells dataset described in this note is limited to the period between late September 2008, when the current standard was introduced, and the point of data collection in January 2015.

To construct our spells dataset for English Wikipedia, we parsed both the XML log data and SQL page information dumps. Creating spell data required that we track all available records of page protection and unprotection events, page deletions that end protection spells, and events when protected pages were moved. Protection settings often "follow" a page when it is moved. In our dataset, move events are coded as the end of one spell at the source of the move and the start of a new spell at the destination. We have released our full source code to document and reproduce our process.[6] Our code consists of more than 300 lines of Python used to parse XML and SQL dumps and more than 650 lines of R to create spells.

Our dataset remains necessarily limited, incomplete, and subject to missing and incorrect input data. In the case of 23,272 spells affecting 12,833 pages, we use SQL dumps to infer protection periods that are "left censored," (i.e., they began before our data collection window [10]). In addition, because protection expirations and page deletions are not noted as protection events in the log, our dataset is missing spells for pages protected before our data collection window if protection expired or if these previously protected pages were deleted within the window. If the protection status of a page protected prior to the window is changed, we report that the earlier protection status existed, but the level and type of prior status is missing. Our dataset does not include "cascading" protection that affects pages that are linked from special protected pages.[7]

[1] https://en.wikipedia.org/wiki/Wikipedia:Protection_policy

[2] https://en.wikipedia.org/wiki/Special:Log

[3] https://en.wikipedia.org/wiki/Wikipedia:User_access_levels

[4] http://www.mediawiki.org/wiki/API:Logevents

[5] http://dumps.wikimedia.org

[6] http://communitydata.cc/wiki-protection

[7] Cascading protection is rare and was applied to only 97 of

	Page Title	Type	Level	Start	End
1	Biology	edit	autoconfirmed	2008-09-29 15:14:44	2008-10-29 15:14:00
2	Biology	edit	autoconfirmed	2008-12-04 03:44:45	2008-12-25 03:44:00
3	Biology	edit	autoconfirmed	2010-03-01 16:20:48	<NA>
4	Biology	move	sysop	2010-03-01 16:20:48	<NA>

Table 1: Example protection spells for one article in our dataset. "<NA>" indicates that the data is censored because the spell was ongoing at the point of data collection.

In other cases, we must make decisions when our two input sources contain contradictory data. We include "bookmarks" in our code (e.g., "BK-X") to help readers to find and inspect these choices. For example, we exclude spells from 8 pages where log and SQL data suggest that a page was protected at contradictory levels (BK-A). We also exclude 4,148 spells whose final protection status in the SQL dumps is directly contradicted by our parsed log data (BK-B). We include 685 additional spells that final state data indicates have ended, but for which the logs do not indicate an end-point (BK-C).

While further development may improve data quality, hand-checking random subsets of contradictory spells has revealed some cases when the log and SQL dumps are simply ambiguous, missing, or incorrect in ways that additional analysis will not be able to fix. For example, at the time of writing, the log clearly records that the last event the disambiguation page "Moist" experienced was an indefinite semi-protection in March 2010,[8] while the page was verifiably editable by unregistered users. We include an "open" spell even though other evidence suggests it ended at some point between 2011 and 2015. Our source code can be modified to handle these cases differently.

Consistent with the WMF Open Access policy, we have published this page protection dataset freely for other researchers under the same license used for all Wikipedia content. Because WMF continues to publish new SQL snapshots of page info – and because other wikis using MediaWiki could be used to generate similar datasets – we have also released all of the software used to create the page protection dataset under the GNU GPLv3.

4. PAGE PROTECTION IN WIKIPEDIA

Our dataset contains 355,532 page protection spells applied to 125,951 unique pages or 0.36% of all pages in English Wikipedia. This counts all of the different "namespaces" in Wikipedia, including pages used for administrative work, discussions, file hosting, and more. Page protection occurs unevenly across these namespaces. For example, the article or "main" namespace accounts for less than 14% of the total pages in the encyclopedia, but 66% (235,458) of all spells.[9] Within the article namespace, 0.67% of pages have been protected during our data collection window.

The dataset suggests that the total number of pages protected at any given point has remained relatively stable increasing at about 1,000 new protected pages per year from 45,000 in 2008. Despite the relatively stable number of protected pages, there is a great deal of dynamism and variation within the population of spells. Among the page protections that start and end within the period covered by our dataset, protection spells are as short as a few seconds and as long as 6 years with a median of 14 days ($\bar{x}$: 123, σ: 292). At the point of data collection, 27% of the spells in the dataset were ongoing. Page protection shifts frequently and is distributed unevenly, with 96% of pages that have ever been protected undergoing multiple protection spells. Among pages that have experienced at least one protection, the median number of protections is 2 ($\bar{x}$: 2.8, σ: 3.5) with the talk page of one administrator, "NawlinWiki" – a target of persistent vandalism – experiencing a total of 797 protection spells during the window.

	Protected	Min	Median	Mean	Max
Unreg.	Yes	0	0	0.6	1,851
only	No	0	0.2	13	151,200
All	Yes	0	0.48	410	483,839
editors	No	0	1.1	110	302,400

Table 2: Edits per week by unregistered editors and all editors for pages in our dataset while protected and unprotected.

5. IMPACT OF PAGE PROTECTION

Page protection has dramatic effects on the dynamics of editing. Comparing editing activity by unregistered editors within pages that are protected and unprotected illustrates the impact of page protection and also provides a validity check for our dataset. The top section of Table 2 compares edits per week from unregistered editors for pages included in our dataset during periods when they are protected and unprotected. Unregistered editors make edits during only 6% of protection spells. By contrast, unregistered editors edit in 65% of the unprotected spells for the same pages. Edits from unregistered editors while pages are protected reflect the irregularities described above. The bottom section of Table 2 compares total edits per week for pages included in our dataset. Both a t-test for a difference of means ($t = 5.53$, p: 0) and a non-parametric Kolmogorov-Smirnov test for whether the two groups are drawn from the same underlying distribution (D: 0.15, p: 0) indicate that these differences are statistically significant.

Research analyzing the patterns and structure of editing should take page protection into account because protection systematically prevents some editors from contributing. For example, Keegan et al. published a series of influential papers comparing the structure of editing on breaking news articles about airplane crashes with articles written about historical accidents (e.g., [6]). Although the large majority of the contributions to the 249 articles studied by Kee-

106,103 ongoing spells at the point of data collection.
[8]https://en.wikipedia.org/wiki/Special:Log?page=Moist
[9]https://en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&oldid=663786511

	M	$\bar{x}$	Range
Mean Views (All)	5.23	5.24	[3.89,6.06]
Mean Views (Protected)	207	195	[71.8,294.07]
% Views (Protected)	14.3	13.6	[6.66,21.46]
% Articles (Protected)	0.36	0.37	[0.33,0.57]
% in Top 1000 (Protected)	34.1	34.2	[27.3,42]

Table 3: Summary statistics for measures of views during 168 one-hour periods in December 2013. Columns are included for the median (M), mean ($\bar{x}$), and range.

gan et al. were made before the window of our dataset, we found protection spells for 12 articles within our window. Although 93 articles were classified as breaking news, 9 of the 12 protected articles were breaking news events, more than we would expect by chance (χ^2: 6.04, df: 1, p: 0.01).

Accounting for page protection should improve the precision of Keegan and colleagues' analysis of editing dynamics around breaking news. The fact that breaking news articles are more likely to be protected shapes the relationship between breaking news status and key variables like the distribution of editors across articles, the "tempo" of activity, and back-and-forth patterns of editing. Likewise, studies of the relationship between editing activity and article quality as measured by featured article status (e.g., [9]) should address the impact of page protection since all featured articles are protected when they are posted on the Wikipedia front-page.

Protected pages are also disproportionately viewed by Wikipedia readers. Table 3 shows data on viewership during during the first week of December 2013 (168 hour-long periods) for pages that were protected for at least part of each hour and for all viewed pages. On average, protected articles were only 0.37% of the articles viewed, but received 14% of views. The median protected article received 26 views ($\bar{x}$: 195, σ: 6,363) while the median unprotected article received 1 view ($\bar{x}$: 5, σ: 391). Among the top 1,000 most viewed articles during each of these hours, an average of 342 were protected.

Research that exploits the role of page views in shaping contributions is at risk of systematically underestimating the strength of this relationship unless it accounts for page protection. For example, Gorbatai [3] shows how viewership drives edits by newcomers, which in turn drives engagement by established editors, leading to improved article quality. Because highly-viewed articles are more likely to be protected, the association between edits from newcomers and edits from established editors would likely be even stronger if page protection were considered. These findings also illustrate how page protection may undermine a key mechanism driving the improvement of highly-viewed articles.

6. CONCLUSIONS

Page protection shapes the structure and experience of collaborative work in wikis. We have demonstrated how page protection impacts the relationship between article views and edits. Studies of wiki work and Wikipedia governance might use this dataset to investigate how Wikipedians manage the boundaries of collaboration. Granular, precise data about page protection also facilitates studies of related phenomena. For example, sudden changes in the number of edits to a page brought about by a protection event create a discontinuity that could be used to estimate the impact of blocking unregistered contributors.

Future work can determine how best to incorporate a detailed understanding of page protection into existing knowledge about Wikipedia and other wikis. Page protection also merits more systematic analysis in its own right. Such analysis lies beyond the scope of this paper, but we hope that others will use our dataset and software toward this end. As researchers begin to consider protection, we invite contributions to our code that address the limitations we have described and other issues of which we are unaware.

7. REFERENCES

[1] J. Antin and C. Cheshire. Readers are not free-riders. In *Proc. CSCW '10*, pages 127–130. ACM Press, 2010.

[2] S. L. Bryant, A. Forte, and A. Bruckman. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proc. SIGGROUP '05*, pages 1–10. ACM Press, 2005.

[3] A. Gorbatai. Aligning Collective Production with Demand : Evidence from Wikipedia. Available at SSRN 1949327, 2011.

[4] B. M. Hill and A. Shaw. Consider the redirect: A missing dimension of Wikipedia research. In *Proc. OpenSym '14*, pages 28:1–28:4. ACM Press, 2014.

[5] D. Jemielniak. *Common Knowledge?: An Ethnography of Wikipedia*. Stanford University Press, Stanford, California, May 2014.

[6] B. Keegan, D. Gergle, and N. Contractor. Hot Off the Wiki Structures and Dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral Scientist*, 57(5):595–622, May 2013.

[7] D. W. McDonald, S. Javanmardi, and M. Zachry. Finding patterns in behavioral observations by automatically labeling forms of wikiwork in Barnstars. In *Proc. WikiSym '11*, pages 15–24. ACM Press, 2011.

[8] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. SIGGROUP '07*, pages 259–268. ACM Press, 2007.

[9] S. Ransbotham and G. C. Kane. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly*, 35(3):613, 2011.

[10] J. D. Singer and J. B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, New York, 2003.

[11] F. B. Viegas, M. Wattenberg, M. M. McKeon, and D. Schuler. The hidden order of Wikipedia. In *Online Communities and Social Computing*, HCII, pages 445–454. Springer-Verlag, Berlin, 2007.