Consider the Redirect: A Missing Dimension of Wikipedia Research

Benjamin Mako Hill
University of Washington
Department of Communication
Seattle, WA 98195
makohill@uw.edu

Aaron Shaw
Northwestern University
Department of Communication Studies
Evanston, IL 60208
aaronshaw@northwestern.edu

ABSTRACT

Redirects are special pages in wikis that silently transport visitors to other pages. Although redirects make up a majority of all article pages in English Wikipedia, they have attracted very little attention and are rarely taken into account by researchers. This note describes redirects and illustrates why they play an important role in shaping activity in Wikipedia. We also present a novel longitudinal dataset of redirects for English Wikipedia and the software used to produce it. Using this dataset, we revisit several important published findings about Wikipedia to show that accounting for redirects can have important effects on research.

Categories and Subject Descriptors

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

General Terms

Wikipedia, Peer Production, Redirects, Wikis

1. INTRODUCTION

A feature of most wikis, "redirects" are special pages that silently transport visitors to other pages. In Wikipedia, the only indication that one has visited a redirect is that the page title and the URL in the browser are different and a very small hyperlinked message appears near the article title (see Figure 1). Clicking on this link will take the user to the redirect page itself. Redirects in Wikipedia are normal pages that begin with "`#redirect [[Target]]`" where "`Target`" is the page to which visitors will be redirected. Although redirect pages can contain extensive text, their content is almost never viewed and very rarely edited. Despite their near-invisibility, redirects play an important role in shaping activity in Wikipedia. Redirects are a majority of all article pages in English Wikipedia and are viewed millions of times each month. They represent a central form of the encyclopedia's "hidden order" [7] and contribute to wikis' usability and user experience.

OpenSym '14 , Aug 27-29 2014, Berlin, Germany
ACM 978-1-4503-3016-9/14/08.
http://dx.doi.org/10.1145/2641580.2641616

Figure 1: An example of a redirect in English Wikipedia. Note the small redirect notice below the title.

That said, redirects have attracted very little attention from researchers studying Wikipedia and are, with rare exceptions (e.g., [3, 8]) rarely discussed explicitly in the analysis of Wikipedia data. In this note, we make several contributions: First, we introduce a longitudinal database that makes it easier to study redirects in English Wikipedia over time and use this database to characterize the enormous volume of activity around redirects. Then, we use the database to illustrate the importance of considering redirects in two relationships of central interest to many researchers: (1) the distribution of edits over articles and (2) the relationship between views and edits. We conclude with guidance for how researchers should account for redirects in future work.

2. DESCRIBING REDIRECTS

Redirects constitute a key piece of Wikipedia's invisible infrastructure that supports the broader usability of the site. They are also an example of how some of the least visible activities and elements of wikis lend structure and salience to visitors' experience.

Redirects connect many types of pages. For example, redirects link alternative spellings or transliterations (e.g., from "Mohammed" → "Muhammad"), common abbreviations (e.g., "NYC" → "New York City"), translations (e.g., "Nueva York" → "New York City") or common misspellings (e.g., "Barrack Obama" → "Barack Obama"). Plural nouns will often redirect to the singular form (e.g., "Frogs" → "Frog"). Like many other wiki platforms, MediaWiki (the software that runs Wikipedia and other Wikimedia Foundation (WMF) Wikis) treats articles with differences in case and spacing as distinct. As a result, many redirects exist between articles that only differ in capitalization and white-space (e.g., "Newyorkcity" and "New york city" → "New York City").

Redirects perform several critical functions within Wikipedia [3]. They facilitate easy hyperlinking by creating targets from alternative names and spellings. They facilitate moving pages and other maintenance by ensuring that the flow of traffic throughout the wiki is uninterrupted. They provide common shortcuts for discussing content or policies. They help search engines – including Wikipedia's internal search engine – locate the correct information for a particular search term. They also contribute to the options in Wikipedia's auto-complete search box.

The creation and maintenance of redirects are examples of Wikipedia's hidden order because, like other pages in Wikipedia, redirects are products of collaborative effort. Wikipedia editors create redirects and then update them as pages move and redirects' targets change. Controversial redirects are discussed in a "Redirects for Discussion" page,[1] where both likely outcomes (i.e., decisions to delete or to change the target of a redirect) will be invisible. When an article is renamed, MediaWiki automatically creates a redirect in place of the old article so that existing hyperlinks still function. Like other pages, redirects are placed in categories, discussed on talk pages, and tagged with templates.

Similarly invisible aspects of Wikipedia and work in wiki communities have attracted scholarly attention. Early research into Wikipedia treated the encyclopedia's "hidden order" as the behind the scenes coordination and discussion that editors engaged in to create the public-facing pages that most visitors to the site interact with [7, 6]. Related work examined invisible and peripheral forms of participation in Wikipedia and other online communities, demonstrating that these forms of labor are nonetheless legitimate modes of collaboration (e.g., [1, 2]). More recent studies have examined how different kinds of labor can give rise to distinct "social roles" and types of "wikiwork" [5, 9]. Redirects offer another example of how invisible and seemingly peripheral modes of participation in social computing systems perform valuable socio-technical functions.

3. LONGITUDINAL REDIRECT DATASET

One challenge with studying redirects in Wikipedia is that common sources of Wikipedia data treat redirects differently. For example, many requests to the MediaWiki API[2] will automatically follow redirects. Although the API indicates the presence of a redirect in metadata, it can effectively hide redirects from researchers. Other widely used data sources, like WMF's database dumps and page view data,[3] present redirects as normal pages, inviting different methodological and substantive problems [3]. The WMF has, at least since 2013, published a database of redirects as an SQL "dump" file for every wiki they host. These data reflect a cross-sectional snapshot of the redirect network within the wiki at the time that the dump is produced. Unfortunately, cross-sectional data on redirects is insufficient for many purposes because Wikipedia's redirect network changes over time as editors make changes to redirects and their targets.

For example, the Wikipedia article on "NA" was initially created as a disambiguation page in January, 2004, edited 26 times, and then turned into a redirect to "Na" in February, 2006 when the "NA" and "Na" disambiguation pages were merged. Similarly, redirects can be turned into normal articles. For example, articles about characters in popular television series will often begin as redirects to the article about the show. When coverage of characters in the show's article grows, the redirects can be turned into stand-alone articles. Even within articles that are stable as redirects, targets can change over time (see Table 1 for an example).

Comprehensive dynamic data on redirects in Wikipedia is difficult to create because it requires looking within the text of the revisions of all articles. This full text history is currently several terabytes of uncompressed XML data and can require hundreds of hours of computer time to process. Using a research computing cluster at the Harvard MIT Data Center (HMDC), we parsed the full text revision history of English Wikipedia published in October, 2012, to create a list of all revisions that redirect. For each redirect, we recorded the revision timestamp as well as the page id, title, and the target.

From this process, we created a dataset of *redirect spells*. Each "spell" includes the title of the redirect or (i.e., the page that is being redirected from) the target of the redirect (i.e., the page being redirected to) and the start and end timestamps for the spell. This dataset includes 9,277,563 observations. Many redirect spells (67%) did not end (i.e., pages remained redirects at the time that the data was collected) so the "end" date is censored and marked as missing for these spells. Examples of spells for one page are shown in Table 1. In the example spells in the table, a single redirect was (1) created, (2) changed to redirect to a different target, and finally (3) changed back to the first target until at least the point of data collection.

We have published our dataset freely for other researchers under the same license used for all Wikipedia content. Because WMF continues to publish new dumps of Wikipedia – and because all other MediaWiki wikis use redirects – we have also made all of our software available for free under the GNU GPLv3. Both code and dataset are available on our website.[4]

4. REDIRECTS IN WIKIPEDIA

Our dataset contains details on 6,201,873 redirect pages representing 22% of pages hosted in English Wikipedia. This includes all of Wikipedia's "namespaces" including pages for administrative work, discussion, and file hosting where redirects are not widely used. The vast majority of redirects (5,318,869 or 86%) are in the public-facing article or "main" namespace. Within the article namespace, a majority of pages (55%) are redirects. As shown in Figure 2, the number of redirects in Wikipedia has grown in a nearly linear pattern over time.

The dataset suggests that redirects are indeed dynamic. Redirect spells last as little as a fraction of a second and as long as 10 years with a mean of 363 days (Median: 122, SD:

[1] http://enwp.org/WP:RFD
[2] https://www.mediawiki.org/wiki/API
[3] http://dumps.wikimedia.org/
[4] http://networkcollectiv.es/wiki-redirects/

	Page ID	Page Title	Target	Start	End
1	4151445	Kirtland Perky	Kirtland I. Perky	2006-02-21 17:18:48	2010-10-19 06:14:26
2	4151445	Kirtland Perky	Kirtland Irving Perky	2010-10-19 06:14:26	2011-06-03 05:11:13
3	4151445	Kirtland Perky	Kirtland I. Perky	2011-06-03 05:11:13	<NA>

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

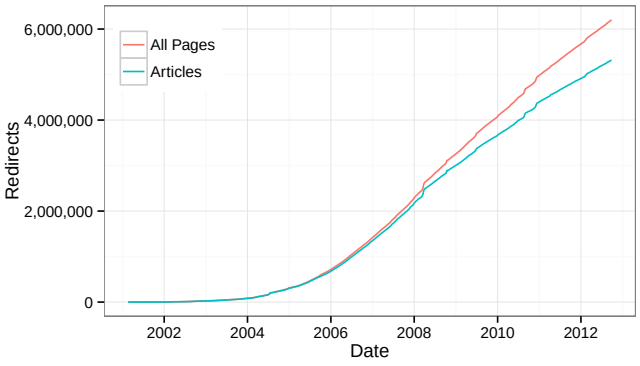

Figure 2: Count of redirects in English Wikipedia over time for all pages and within the article namespace only.

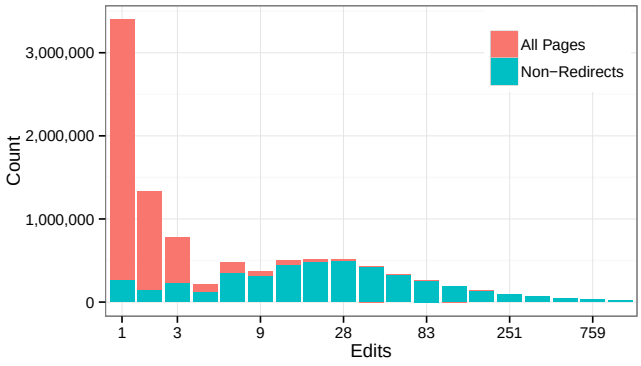

Figure 3: Histograms of pages in the article namespace based on number of edits on a log scale. The red histogram includes all pages in the article namespace ($N = 9,729,989$). The blue graphs includes only pages that were not redirects at the point of data collection ($N = 5,327,561$).

512). As we have noted, 67% of spells in the dataset were ongoing at the point of data collection. The dynamic dataset also shows that redirects are frequently shifting. 3.2% of articles (1.4% of all pages) were once redirects but were not at the time of our data collection. 24% of articles that have redirected once have experienced multiple spells and one administrative page experienced 713 different spells.

5. IMPACT OF REDIRECTS

The fact that such a large proportion of pages in English Wikipedia are redirects has important implications for Wikipedia research. That said, little previous research has explicitly taken redirects into account. For example, work examining the distribution of edits across different types of pages (e.g., [4]) should exclude or adjust for redirects because they are substantively different than other pages and systematically less visible to editors. Likewise, research analyzing the distribution of links or edits over pages (e.g., [8]) should adjust for the presence of redirects because they are systematically less likely to incorporate links or be edited.

In general, since redirects are both unfamiliar and invisible to the vast majority of wiki readers, it may make sense to either remove redirects entirely from analyses or to collapse activity on redirects with activity on their target pages. While this can be difficult for longitudinal analyses (due to the dynamic nature of redirects), it yields substantively different results under a wide variety of circumstances. We illustrate this in Figure 3 where we plot the distribution of edits per article as of October, 2012 with and without redirects. Because most articles are redirects and most redirects are edited very infrequently, the two distributions suggest radically different distributions of edits across pages. This has an important impact both on the methods researchers use to model Wikipedia activity and on our substantive understanding of the ways editors distribute their work across pages.

A growing body of studies using Wikipedia viewership data presents another area in wiki research where redirects are important but overlooked. Because viewers don't see redirects, viewing a redirect is substantively different from viewing a normal page. For example, if a user visits the article on "Seattle, Washington", this will be recorded as a view to the redirect even though the target article "Seattle" is displayed. In this sense, views of redirects will tend to be overcounted while views of target articles will tend to be undercounted. If a visitor subsequently presses the edit button on the top of the page, they will edit the target.

In a highly-cited article, Priedhorsky et al. combine Wikipedia view data with edit data and do not mention accounting for redirects [6]. Although not a central finding in their paper, the authors note with surprise that they find a weak correlation between edits and views. Because redirects are edited infrequently but "viewed" as often as millions of times per month each, redirects may be contributing to the surprisingly low correlation between edits and views noted by Priedhorsky et al. and others.

We use the closest full month of public viewership data to the period studied by Priedhorsky et al. (January, 2008) to revisit their finding with redirects in mind. First, we create cumulative tabulations of views (over January, 2008) and edits (over all time) for every page in English Wikipedia

Type of Views	Correlation statistic	
	Pearson (log)	Spearman
Articles		
Raw	0.67	0.64
Redirect-adjusted	0.81	0.79
All pages		
Raw	0.61	0.5
Redirect-adjusted	0.74	0.65

Table 2: Correlation between edits and views for pages in English Wikipedia in the article namespace ($n = 9,729,989$) and for all pages ($n = 28,397,622$). Raw edits reflects the views to articles as recorded in public page view data. Adjusted edits "shift" views from redirects to their targets. Pearson correlations are of log transformed edits and views.

with at least one edit. We then calculate the proportion of each month that each page redirected to each of its targets. Next, we multiply these proportions with counts of views for the redirects to create an estimate of the views made to pages while they were redirects. Finally, we subtract these products from the views tabulated for each redirect page and add them to the redirects' targets. For example, if the article named "Seattle, Washington" was viewed 100 times in January, 2008 and was a stable redirect to the article named "Seattle" for the entire month, we would shift all of the redirect's views to "Seattle." However, if the redirect was a stand-alone article for two weeks out of the month, we would shift only half of the views.

Table 2 shows that adjusting views for redirects substantially increases the correlation between edits and views. The Pearson correlation between log views in January, 2008 and log edits made to articles through that period increases from $\rho = 0.67$ to $\rho = 0.81$ when we consider redirects. We see a similar pattern and magnitude of increase in the strength of correlations using the full dataset of pages in Wikipedia and the non-parametric rank-based Spearman correlation.

6. CONCLUSIONS

Redirects constitute an important component of the structure of wikis. Researchers must reckon with redirects more carefully and explicitly in order to accurately represent the structure of wikis as well as the flow of work and attention of contributors and viewers. As a hidden but central aspect of Wikipedia's infrastructure, redirects deserve deeper and more sustained attention in their own right. For example, redirects are an untapped source of dynamic network data within Wikipedia. They are also a distinct form of contribution and viewership behavior that previous approaches have not explored in depth.

Future research can utilize the database of redirects we have created and the code used to generate it. For example, these data could be used in network analyses involving relational models of pages, articles, edits, or views to account for redirects. Likewise, studies considering the dynamic relationship between viewership and editorship will gain precision by incorporating redirects into their analysis. In addition, the large body of work mining Wikipedia's link and knowledge structure may uncover novel relationships and dynamics by considering redirects. This list is not exhaustive, but suggests situations in which including redirects in quantitative analysis may alter the magnitude, if not the direction, of research findings. Above all, we hope this note inspires future research to handle redirects in a more explicit and consistent manner. Given how little-analyzed and poorly-understood redirects have been, further work will be necessary to discover the places where redirects enhance our understanding of Wikipedia and other wikis.

7. REFERENCES

[1] J. Antin and C. Cheshire. Readers are not free-riders. *Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10)*, pages 127–130, 2010.

[2] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, GROUP '05, page 1–10, New York, NY, USA, 2005. ACM.

[3] B. J. Hecht. *The Mining and Application of Diverse Cultural Perspectives in User-Generated Content*. Ph.D. dissertation, Northwestern University, Evanston, IL, 2013.

[4] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in wikipedia. In *Proc. CHI '07*, 2007.

[5] D. W. McDonald, S. Javanmardi, and M. Zachry. Finding patterns in behavioral observations by automatically labeling forms of wikiwork in barnstars. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 15–24, New York, NY, USA, 2011. ACM.

[6] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, page 259–268, New York, NY, USA, 2007. ACM.

[7] F. B. Viégas, M. Wattenberg, M. M. McKeon, and D. Schuler. The hidden order of wikipedia. In *Online Communities and Social Computing*, HCII, page 445–454. Springer-Verlag, Berlin, Heidelberg, 2007.

[8] J. Voss. Measuring wikipedia. In *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.

[9] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, page 122–129, New York, NY, USA, 2011. ACM.