Characterizing “Permanently Dead" Links on Wikipedia

Anish Nyayachavadi Jingyuan Zhu Harsha V. Madhyastha
University of Michigan

Abstract

It is common for a web page to include links which help visitors discover related pages on other sites. When a link ceases to work (e.g., because the page that it is pointing to either no longer exists or has been moved), users could rely on an archived copy of the linked page. However, due to the incompleteness of web archives, a sizeable fraction of dead links have no archived copies.

We study this problem in the context of Wikipedia. Broken external references on Wikipedia which lack archived copies are marked as “permanently dead". But, we find this term to be a misnomer, as many previously dysfunctional links work fine today. For links which do not work, it is rarely the case that no archived copies exist. Instead, we find that the current policy for determining which archived copies for an URL are not erroneous is too conservative, and many URLs are archived for the first time only after they no longer work. We discuss the implications of our findings for Wikipedia and the web at large.

CCS Concepts

• **Information systems → World Wide Web**; **Digital libraries and archives**.

Keywords

Link rot, web archives

ACM Reference Format:
Anish Nyayachavadi, Jingyuan Zhu, and Harsha V. Madhyastha. 2022. Characterizing “Permanently Dead" Links on Wikipedia. In *ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3517745.3561451

1 Introduction

The authors of a web page often include links to other pages on the web, in order to point the page's visitors to related information and services. But, these links cease to work over time, a problem commonly referred to as “link rot" [10, 25, 33]. When users attempt to visit dysfunctional links, they might encounter a variety of errors such as failed DNS lookups, connection timeouts, or “page not found" responses. These errors occur either because the site hosting the linked page has been abandoned or that page has been deleted or moved.

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10…$15.00
https://doi.org/10.1145/3517745.3561451

In this paper, we focus on link rot in Wikipedia's articles. Wikipedia makes for an important case study for two reasons. First, while broken links may be a mere annoyance elsewhere on the web, their presence in a Wikipedia article significantly hampers the verifiability of the article's content. Hundreds of thousands of users have spent time and effort in including appropriate citations across all of Wikipedia. But, link rot is putting a lot of this work to waste. Second, unlike elsewhere on the web where one can only identify which links are broken, Wikipedia enables users to edit articles and patch broken links. As a result, when a reference in any article no longer works, software such as the InternetArchiveBot [27] can augment the reference with a link to an archived copy of the referenced page.

However, a fundamental limitation of relying on archived page copies to fix dead links is that, given the scale of the web, even the largest archive that exists today (the Internet Archive) is incomplete. For a large fraction of publicly accessible URLs, either no archived copies exist or all copies were captured only after the URL stopped functioning. On Wikipedia, a link to any URL which lacks non-erroneous archived copies is marked “permanently dead". Such broken references are particularly undesirable because the content that enables verifiability is accessible neither at the original link nor via a web archive.

To understand the factors that cause links to end up being marked as permanently dead, we analyze a random sample of 10,000 such links from the English Wikipedia. We make three broad observations, all of which have significant implications for how to deal with broken links and how to reduce the number of links that have no archived copies. While we make these observations on Wikipedia, the implications are applicable to any page on the web.

Dead links do not remain broken forever. First, we found that 3% of the links which have been marked as permanently dead are in fact functional today. This is not because some links have been erroneously tagged as permanently dead. Rather, in cases where a URL is rendered dysfunctional because the page it points to has been moved, that URL might work again in the future when the site hosting the page adds a redirection to the page's new URL. Therefore, the term “permanently dead" is not always appropriate for broken links which have no archived copies.

Usable archived copies are often unused. Second, while most permanently dead links indeed do not work today, we found that there exist usable archived copies for over 15% of these links. InternetArchiveBot – the dominant software for finding and patching broken links on Wikipedia – marked such links as permanently dead because of two reasons. To operate efficiently at scale, the bot assumes that a link was never archived if its attempt to lookup archived copies for that link does not complete in a timely manner. In addition, because redirections on the web are often erroneous (e.g., the old URL for a news article might redirect to the news site's

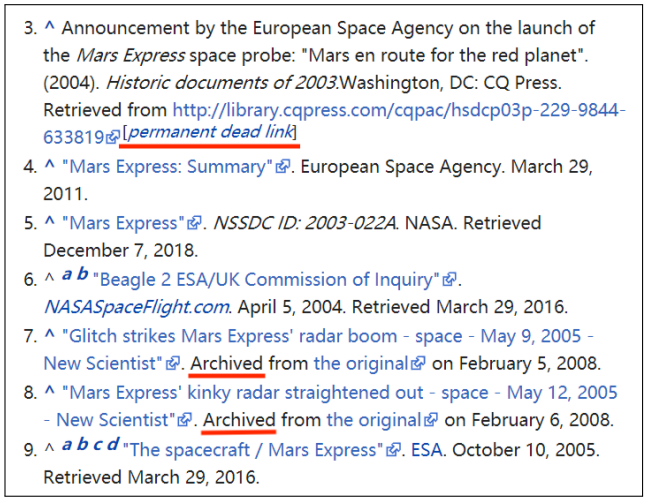


Figure 1: Example of broken links augmented by InternetArchiveBot on Wikipedia.

homepage), it conservatively links to a page's archived copy only if no redirections were encountered when that copy was crawled. Instead, by carefully distinguishing erroneous redirections from non-erroneous ones, we find that 5% of broken links which are currently tagged as permanently dead could be patched with links to their archived copies.

Links must be sanity-checked and archived when posted. When no usable archived copy exists on the Internet Archive for a broken link, we see that the reasons are primarily two-fold. In cases where users made typos when posting links, these links were never functional to begin with; it was inevitable that they would end up as permanently dead. More importantly, for almost half of the permanent dead links, the Internet Archive first attempted to archive them only several months or even a few years after those links were added to Wikipedia; by then, they had stopped functioning. Both of these issues can be avoided if, whenever a link is posted, the liveness of the link is confirmed and an archived copy is captured soon thereafter.

2 Setting and Data

We begin by describing the context for our work, the sources of data that we rely upon for our analysis, and the goals of our study.

2.1 Coping with link rot on Wikipedia

All across the web, there exists the risk that links to pages on other sites may stop working a few years after those links are created. However, while one can write software to identify the broken links on any page, only a website's provider can rewrite the pages on that site to patch these links. In contrast, since any user can edit any article on Wikipedia, there exist several bots which attempt to not only find broken external links but also augment them with pointers to the corresponding archived copies.

The InternetArchiveBot (IABot) [27], which is administered by the Internet Archive, is the most well-known among all such bots on Wikipedia. Whenever it scans any article on Wikipedia, IABot extracts all outgoing links and tests which of them are broken. IABot determines that a URL is broken if its HTTP GET request for that

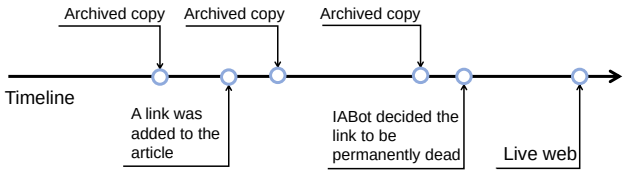


Figure 2: Timeline depicting different points of interest in our analysis of any link which has been marked permanently dead. Note that some, if not all, of the archived copies for a URL may have been captured after that URL stopped functioning.

URL does not result in a 200 status code response (after potential redirections). Note that IABot's goal is not to comprehensively detect all broken links, but to patch links that are definitely broken. It augments every broken link with a reference to an archived copy of that link (hosted either on the Internet Archive's Wayback Machine [16] or on one of more than 20 other web archives [28]). Since every web archive captures many snapshots of every URL, IABot augments a broken link on any article with that archived copy for the link which was captured closest to when the link was added to the article. As an example, Figure 1 shows how IABot has added links to archived copies for references 8 and 9 on the Wikipedia article https://en.wikipedia.org/wiki/Mars_Express.

2.2 Permanently dead links

On Wikipedia, a broken link for which no archived copy exists is tagged as a *"permanent dead link"*. Reference 3 in Figure 1 is an example of IABot tagging a link as such. When we conducted our study in March 2022, there were over 180,000 articles on the English Wikipedia which included links marked as permanently dead [31]. We crawled all of those articles and found 290,669 unique URLs tagged as permanently dead in total.

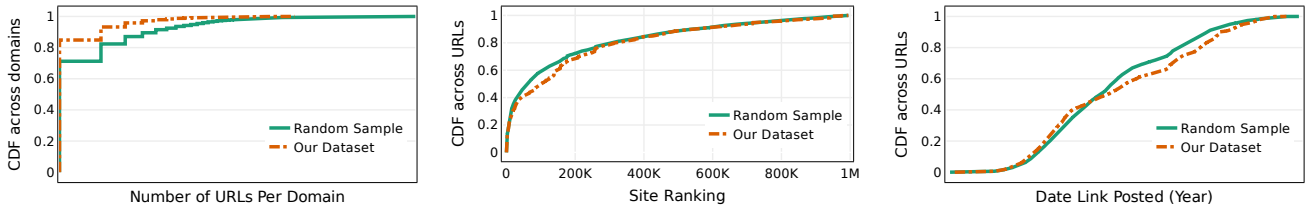
2.3 Goals of our study

Permanently dead links particularly hamper user experience on Wikipedia because, not only does the original link no longer work, but there is no valid archived copy of that link which a user could refer to. As a result, readers of any article which includes a permanently dead link can either no longer verify the claim that this link was meant to support or they miss out on additional information/services that the link originally led to.

Therefore, in this paper, we study links which have been marked as permanently dead on Wikipedia. We examine each such link from the following perspectives.

- What is the status of the link on the web today?
- What archived copies, if any, existed for the link before it was marked permanently dead? If any archived copies existed, were they all captured after the link stopped functioning?
- When was the first archived copy captured relative to when the link was posted on Wikipedia?
- Are there no valid archived copies for the entire site on which the linked page is hosted, or was only this specific URL not archived out of all pages on that site?

To answer these questions, Figure 2 shows the events of interest for any particular permanently dead link.



(a) (b) (c)

Figure 3: For 10,000 permanently dead links that we analyze, distribution along three dimensions: (a) Number of URLs per domain, (b) Site ranking, and and (c) Date when the link was added to Wikipedia. Note logscale on x-axis of (a).

2.4 Dataset

Over the course of March 2022, we collected three types of data from Wikipedia, the Internet Archive's Wayback Machine [16], and the live web.

- First, we fetched and parsed the current version of 10,000 articles which contain links that have been marked permanently dead. The English Wikipedia's collection of articles with permanently dead links [31] lists such articles in an alphabetical order of the article's title. We crawled the first 10,000 articles in this alphabetical order. In total, we found roughly 17,000 unique URLs that had been marked as permanently dead.
- Second, we fetched the entire edit history of each article. We use an article's edit history to identify three pieces of information for every link which is marked permanently dead: 1) the date on which the link was originally added to the article, 2) the date on which the link was marked permanently dead, and 3) the username which marked the link as permanently dead; any Wikipedia user can annotate any link as a "permanent dead link", and every bot that is approved to run on Wikipedia has an associated username too. Based on this information, we randomly sampled 10,000 URLs which have been marked as permanently dead by IABot. We focus on links that have been marked as permanently dead by IABot for two reasons: a) it is responsible for the vast majority of broken links in our data, and b) its code is open-source [3], which enables us to reason about its operation.
- Finally, for each of the selected 10,000 broken URLs, we queried the Internet Archive's Wayback Machine for its archived copies of that URL. For every archived copy, we logged the timestamp at which it was captured and the initial HTTP status code associated with that copy; for any URL, we use the term initial (final) status code to refer to the status code seen in response to a HTTP GET request for that URL prior to (after) all redirections.

The 10,000 permanent dead links in our dataset are spread across 3,521 domains and 3,940 hostnames. We extract the hostname from any particular URL as the portion of the URL between the protocol (i.e., "http://" or "https://") and the first '/' thereafter. We map any URL's hostname to its corresponding domain using data from the Public Suffix List [21]. Figure 3(a) shows that the distribution of URLs across domains is heavy-tailed; over 70% of domains contribute a single URL, whereas a small number of domains contribute more than 100 URLs each. Figure 3(b) shows that the URLs are spread across sites from a wide range of Alexa rankings.

Moreover, the time at which these links have been posted to Wikipedia spans the last 15 years; the distribution in Figure 3(c) matches the rate at which new articles have been added over time to the English Wikipedia [32]. 40% of these broken links were posted after 2015 and 20% were posted after 2017, which shows that many links become dysfunctional even a few years after they are posted.

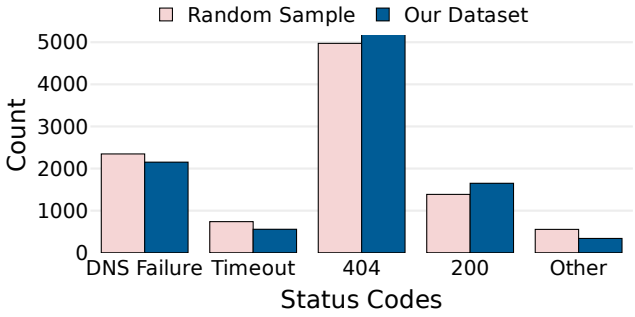


Figure 4: Breakdown of permanently dead links with respect to the outcome when we attempt to fetch them on the live web.

Representativeness of dataset. Later, in September 2022, we crawled all articles on English Wikipedia in which at least one link has been marked permanently dead [31]. Out of all permanent dead links that appear in these articles, we selected 10,000 at random. We found the distributions in Figures 3 and 4 to be largely identical for this random sample and our dataset of 10,000 links.

3 Are permanently dead links indeed dead?

We begin our analysis by examining the current status on the live web of each of the 10,000 URLs in our dataset. We issued a HTTP GET request for every URL and noted the outcome. We classify each outcome into one of five categories:

1. *DNS Failure*: DNS resolution for the hostname in the URL returned an error.
2. *Timeout*: TCP or TLS connection setup timed out.
3. *404*: The final status code was 404 (Not Found).
4. *200*: The final status code was 200 (OK).
5. *Other*: The final status code was neither 404 nor 200, e.g., 503 (Service Unavailable).

As expected, Figure 4 shows that the vast majority (over 70%) of the 10,000 URLs result in either a failed DNS lookup or a 404 response, both of which indicate that those URLs indeed do not work today. A DNS failure is symptomatic of an entire site or subdomain within a site being no longer available. Whereas, a 404 response indicates that specific page does not exist now. In cases where we see a connection timeout or a HTTP error response other than 404, it is hard to tell whether the URL is dysfunctional, the service is temporarily unavailable, or accesses to the page are being

blocked because of our measurement vantage point (e.g., due to geo-blocking [20]).

However, it is surprising that, for a sizeable fraction of URLs – roughly 16% – we obtained a 200 status code response. For example, IABot marked the URL https://www.baltimoresun.com/news/bs-xpm-1993-12-10-1993344239-story.html as permanently dead on the article https://en.wikipedia.org/wiki/06:21:03:11_Up_Evil in February 2021, but worked just fine in March 2022.

Of course, a 200 status response for an URL does not always mean that URL is still valid. For example, http://www.znaci.net/00003/385.htm returns a 200 status code but the response indicates that it is a parked domain [26]. On the other hand, https://www.baku2017.com/en/sports/basketball-3x3/results redirects to https://www.goalku.com/id/soccer/events, which returns a 200 status code response but is unrelated to basketball.

Of the 1,650 URLs in our dataset which resulted in a final status code of 200, we determined that 305 are not soft-404s by adapting a technique from prior work [8]. Given a URL u to test, we obtain a new URL u' which is identical to u except that the suffix in u following the last occurrence of '/' is replaced by a randomly generated string of 25 characters. Since u' is a randomly generated URL (and hence, invalid), we can infer that u is broken if requests for u and u' redirect to the same URL and that URL is not for a site's login page. We also consider u as broken if the k-shingling based similarity [9] between the text in the final responses for u and u' is over 99%. We do not look for identical responses since multiple requests for even the same URL can yield slightly different responses.

We have found no evidence to suggest that these links were marked permanently dead due to a bug in IABot's code or an error in its operation. When checking any link, IABot determines whether the link is dead by attempting to fetch the link only once. This appears to suffice because, out of all permanent dead links which have at least one archived copy after they were marked permanently dead, we find that the first of these copies is erroneous (i.e., 404, soft-404, etc.) for 95% of links.

We believe that most of the URLs which were found by IABot to be dysfunctional in the past but are functional now correspond to cases where a page's URL has changed; requests for the page's old URL were originally returning an error, but now redirect to the page's new URL. Of the 305 functional URLs, we found that 79% redirect to a different URL before finally returning a 200 'OK' response. One example of such an URL is http://www.fishman.com/artists/steve-henderlong, which was marked dead by IABot on 25 September, 2018 [30]. When we checked this URL in both March and September 2022, it instead redirected to https://www.fishman.com/portfolio_page/steve-henderlong/, the new URL for the same page.

Implications. Our findings show that, when a broken link with no good archived copy is found, it is a misnomer to refer to it as a *"permanent dead link"*. The link might well work again in the future. Based on our conversations with the Internet Archive, they are considering using the term *"presumed dead"* instead. In addition, when any bot tests the links on a page, ones that have previously been marked as dead should be occasionally checked again; they should not always be excluded to maximize efficiency, as IABot currently does [3].

4 What archived copies exist for permanently dead links?

IABot marks a broken link as permanently dead if it finds no archived copy for the link where the initial status code was 200. This, however, does not mean that there are *no* archived copies for that link. To understand the distinction between the two points, for every URL in our dataset, we examine the copies for that URL that existed on the Wayback Machine prior to it being marked permanently dead by IABot. We make two interesting observations from doing so.

4.1 200 status code copies

As expected, most links did not have any archived copies with a 200 status code before they were marked permanently dead. However, we surprisingly found that 11% of them (1,082 links) did. For example, in the second reference on https://en.wikipedia.org/wiki/1956_Tasmanian_state_election, IABot marked the URL http://www.parliament.tas.gov.au/php/Almanac.htm as permanently dead in 2021. But, Wayback Machine had 200 status archived copies for this URL even back in 2002.

The root cause here is that IABot is designed for efficient operation at scale. When IABot finds a dead link in a Wikipedia article, it employs a timeout when querying the Internet Archive's Wayback Availability API [4]. If it receives no response within the timeout period, it assumes that no archived copies exist for the link. Due to this optimization, IABot can sometimes miss 200 status archived copies, even if they exist.

In May 2022, we pointed out to the Internet Archive that the Wayback Machine has 200 status code archived copies for many links that IABot has marked permanently dead on Wikipedia. They then ran WaybackMedic [12] – an alternate bot that the Internet Archive uses to patch Wikipedia's broken references – on all links that had been marked permanently dead. WaybackMedic runs more slowly than IABot and its execution requires manual oversight, but it is more comprehensive in finding usable archived copies. The Internet Archive informed us that WaybackMedic was able to patch 20,080 links that were previously deemed permanently dead.

Note that the existence of a 200 status code archived copy for an URL does not always mean that copy is useful. As mentioned before, a 200 status code response can sometimes be indicative of a soft-404. Identifying broken references on Wikipedia which have been patched with erroneous 200 status archived copies is beyond the scope of this work. Our focus is on links which IABot marked permanently dead, as it was unable to find any valid archived copies.

4.2 300 status code copies

Now, we turn our attention to the 89% of the links in our dataset for which no 200 status code archived copies existed prior to IABot marking them permanently dead. Not all of these correspond to cases where either no archived copy existed on the Wayback Machine before the link was marked permanently dead or all archived copies that did exist were erroneous. Instead, as many as 3,776 links had an archived copy with a 3xx status code, indicating a HTTP redirection.

Since some redirections can be indicative of soft-404s, IABot currently ignores archived copies in which a redirection was observed. However, treating all redirections as suspect is overly pessimistic.

For example, on https://en.wikipedia.org/wiki/100-Mosques-Plan, IABot marked the URL http://www.main-spitze.de/region/floersheim/9204093.htm as a permanent dead link in 2018. But, Wayback Machine has an archived copy from 2014 in which that URL has a non-erroneous redirection to http://www.main-spitze.de/lokales/floersheim/index.htm. Thus, instead of marking this link as permanently dead, IABot could have linked to this archived copy.

Of course, the challenge here is precisely why IABot ignores all archived copies that show a redirection: how to tell which redirections are not erroneous? On the live web, we can test the validity of a redirection by issuing requests to a similar URL, as described earlier (§3). We cannot do the same to test whether an archived redirection for an URL is valid since the server hosting that URL may either no longer exist or have been modified.

Instead, we check whether a historical redirection for a specific URL was not erroneous by confirming whether the URL that it was redirected to was unique, i.e., there weren't other URLs under the same directory (share the same URL prefix until the last '/') which had the same redirection around that time. For each 300 status code archived copy, we compare the target of the redirection to those seen for up to 6 other URLs within 90 days of that copy. We found that 481 of the 3,776 URLs we tested had a non-erroneous 300 status archived copy. Thus, we estimate that roughly 5% of all the permanent dead links in our dataset could have been augmented with links to archived copies with redirections.

Implications. The primary takeaway from our analysis in this section is that a sizeable fraction of links that have been deemed permanently dead could have instead been patched using archived copies available for them. On the one hand, the tradeoff between efficiency in looking up archived copies and coverage in using available copies appears worth revisiting. On the other hand, instead of conservatively ignoring all archived copies in which a redirection was observed, careful cross-examination of archived redirections can help reveal which of them are not erroneous.

5 Why no successful archived copies?

From our discussion thus far, for over 80% of the links in our dataset, the Wayback Machine had no usable archived copy before IABot marked the link as permanently dead. Why did the Internet Archive fail to archive these links when they were still functional? Since we previously observed that 1,082 links in our dataset were erroneously marked as permanently dead by IABot, we restrict our analysis here to the remaining 8,918 links for which no 200 status code archived copies exist. These 8,918 links can be partitioned into two subsets based on whether they have any archived copies on the Wayback Machine: 6,936 links have at least one archived copy, and 1,982 links have none. Next, we analyze these subsets separately.

5.1 Temporal Analysis

First, for every link which has at least one archived copy, we look at the time gap between when it was added to Wikipedia and when the first archived copy for it was captured by the Wayback Machine subsequently. The motivation for doing so is that, since any URL was presumably functional when a user added it to a Wikipedia article, archiving it soon after should ensure that a usable copy exists. If we ignore the 619 permanently dead links which had one or more archived copies prior to when the link was posted, that leaves us with 6,317 links which have at least one archived copy after the link appeared on Wikipedia.

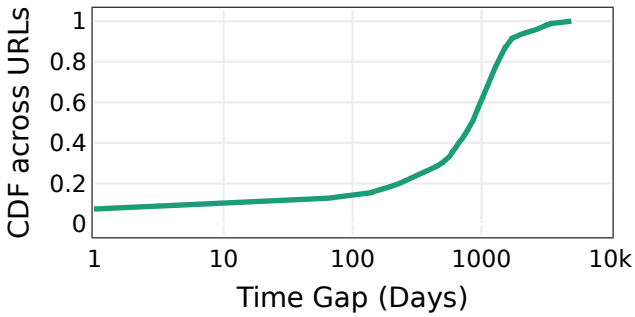


Figure 5: For each permanent dead link that has one or more archived copies on the Wayback Machine, none of which existed at the time the link was posted to Wikipedia, gap between that time and when the first copy for that URL was captured. Note logscale on x-axis.

Figure 5 highlights a key reason for why the Internet Archive failed to capture a working snapshot of any of these links: after a link was added to Wikipedia, the Internet Archive often captured its first copy of that link only several months or years later. During this intermediate period, the URL might transition from a working to an erroneous state. These large gaps are despite the Internet Archive's reliance on the Wikipedia Near Real Time capture service (WNRT) [7] from 2013 to 2018 and on the Wikipedia Eventstream post-2018 [6] to discover and archive links as they are posted to Wikipedia.

For roughly 7% of the 6,317 links (i.e., for 437 URLs), the Internet Archive did manage to capture a copy on the day they appeared on Wikipedia. Yet, we find that 266 of these links surprisingly had an erroneous archived copy even first up! This implies that these URLs were not functional even when a user chose to add them to Wikipedia. Upon manual examination, we find that these cases result from users making typos when posting links. For example, the URL https://www.nj.com/politics/index.ssf/2009/09/nj_gubernatorial_candidates_co.htmlpagewanted=all had a 404 status archived copy on the same day it was posted to Wikipedia because the user who added this link missed adding a '?' or '&' right after html.

Implications. The number of links that have to be marked permanently dead can likely be reduced if the Internet Archive were to more comprehensively archive every URL soon after a link to it is posted on Wikipedia. In addition, when a user is attempting to link to a particular URL from a Wikipedia article, the user needs to be alerted if that URL is dysfunctional.

5.2 Spatial Analysis

Next, we turn to the 1,982 permanently dead links for which *no* archived copies exist on the Wayback Machine. We try to understand why the Internet Archive happened to completely miss these URLs: is the lack of an archived copy due to an isolated coverage gap for a specific page, or due to a larger coverage gap at the granularity of a directory or even an entire hostname? To do so, we query Wayback Machine using its CDX API [5] to find other similar URLs

for which it does have 200 status code archived copies. For each of the 1,982 URLs, we do this once to discover successfully archived URLs which are in the same directory (i.e., share the same prefix until the last '/') and once to discover successfully archived URLs under the same hostname.

From Figure 6, it is evident that most of the coverage gaps involve a specific page, rather than an entire directory or hostname. Of the 1,982 permanently dead links, 749 have no 200 status archived copies at the directory level and 256 have no 200 status archived copies at the hostname level.

We observe that the absence of archived copies for only a specific URL can be attributed to two reasons.

First, many of the URLs which lack archived copies include several query arguments, e.g., http://jhpress.nli.org.il/Default/Scripting/ArticleWin.asp?From=Archive&Source=Page&Skin=TAUHe&BaseHref=DAV/1930/03/19&PageLabelPrint=&EntityId=Ar00305&ViewMode=HTML. In such cases, it is impossible for a web archive to capture all possible URLs on that site; the number of feasible values for some of the query parameters is practically unbounded and different query parameters can appear in any order in the URL.

Second, we again find that some of the URLs which lack archived copies appear to have been mis-typed by the users who added them to a Wikipedia article. For example, the URL http://www.lnr.fr/top-14-orange-histoire-parc-des-princes-paris-26-may-1984-20-08-2004-2-20-10370,10370.html has been marked as permanently dead by IABot on a particular Wikipedia article [29]. But, the user who added this URL mistakenly used the English spelling of the month May in the URL. Replacing "may" in the URL with the French spelling for the month ("mai") makes the URL functional.

We discovered 219 such instances of potential typos using the following methodology. For every URL in our dataset which has no archived copies on the Wayback Machine, we compared the URL to other URLs under the same domain that have been archived. We deem a permanently dead link to potentially be a typo if there exists only one archived URL with an edit distance of exactly 1. If there is no unique archived URL with an edit distance of 1, then the similar URLs typically include a numeric page identifier and it is hard to distinguish a typo versus the omission of a page with a specific identifier.

Implications. Our findings again reinforce the need to alert Wikipedia users when they are attempting to add a link to a dysfunctional URL. Whereas, for URLs which include many query parameters, it might be possible to find archived copies for some of them by either a) identifying which of the parameters specified in the URL have no impact on the page content returned by the server and can, therefore, be ignored, or b) looking for archived URLs which are identical except that they include the query parameters in a different order.

6 Related Work

Link rot. Many prior studies have shown that, when users visit a page created many years ago, they are likely to find that some of the links included on the page do not work [8, 10, 14, 25]. Academic citations and scholarly references have been well-studied given their importance [15, 18, 23]. The commonly used approach to cope with broken links is to rely on archived page copies [33].

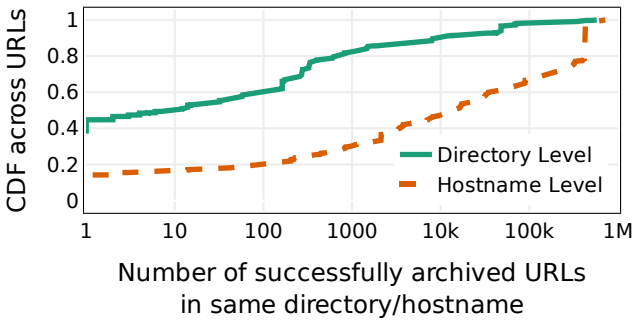


Figure 6: For permanently dead links with no archived copies on the Wayback Machine, number of other URLs – either in the same directory or under the same hostname – for which 200 status code archived copies exist. Note logscale on x-axis.

Coverage of web archives. Prior studies have shown that web archives are not close to archiving all pages on the web [1, 2]. To improve the coverage of web archives, some efforts [13, 19] have attempted to enable the archival of pages that are not publicly available. Many optimizations have also been developed to improve the coverage of web search engine indices [11, 17, 22, 24].

7 Conclusion

Link rot hampers the verifiability of content on Wikipedia, putting to waste the work put into including appropriate references in every article. In this paper, our analysis revealed a number of ways to reduce the number of broken links that cannot be patched with archived copies.

- 3% of the these links are, in fact, not dead links; after previously being dysfunctional, those links work fine now.
- 11% of these links had been marked permanently dead only because of InternetArchiveBot trying to be efficient.
- 5% of permanently dead links could have instead been patched with their archived copies if InternetArchiveBot were to identify and leverage non-erroneous redirections.
- 2% of these links should not even have been permitted onto Wikipedia since users mistakenly added links that never worked.
- The number of broken links that cannot be patched with archived copies can likely be significantly reduced if the practice of capturing a copy of every URL as soon as it is posted on Wikipedia were more comprehensive.

Based on our correspondences with the Internet Archive, they have already applied some of the takeaways from our work. Furthermore, our findings – e.g., the utility of leveraging archived copies corresponding to redirections, the need to alert users when they post dysfunctional links, and the need to archive links as soon as they are posted – are applicable to any site on the web, not only on Wikipedia.

Acknowledgments

We thank Mark Graham and Stephen from the Internet Archive for their comments on earlier drafts of the paper. We also thank Ayush Goel, the anonymous reviewers, and our shepherd Ignacio Castro for their feedback. This work was supported in part by a grant from the Alfred P. Sloan Foundation.

References

[1] Scott G Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. 2011. How much of the web is archived?. In *ACM/IEEE Joint Conference on Digital Libraries*.

[2] Ahmed AlSum, Michele C Weigle, Michael L Nelson, and Herbert Van de Sompel. 2014. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* 14, 3 (2014), 149–166.

[3] internetarchive/internetarchivebot. https://github.com/internetarchive/internetarchivebot.

[4] Wayback Machine APIs. https://archive.org/help/wayback_api.php.

[5] wayback/wayback-cdx-server at master · internetarchive/wayback. https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server.

[6] Wikipedia Eventstream. https://archive.org/details/wikipedia-eventstream.

[7] Wikipedia Near Real Time (from IRC). https://archive.org/details/NO404-WKP.

[8] Ziv Bar-Yossef, Andrei Z Broder, Ravi Kumar, and Andrew Tomkins. 2004. Sic transit gloria telae: Towards an understanding of the web's decay. In *WWW*.

[9] Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Computer networks and ISDN systems* 29, 8-13 (1997), 1157–1166.

[10] Robert P Dellavalle, Eric J Hester, Lauren F Heilig, Amanda L Drake, Jeff W Kuntzman, Marla Graber, and Lisa M Schilling. 2003. Going, going, gone: Lost Internet references. *Science* (2003).

[11] Cristian Duda, Gianni Frey, Donald Kossmann, Reto Matter, and Chong Zhou. 2009. Ajax crawl: Making ajax applications searchable. In *ICDE*.

[12] User:GreenC/WaybackMedic 2.5. https://en.wikipedia.org/wiki/User:GreenC/WaybackMedic_2.5.

[13] Mat Kelly, Michael L Nelson, and Michele C Weigle. 2018. A framework for aggregating private and public web archives. In *ACM/IEEE Joint Conference on Digital Libraries*.

[14] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly context not found: One in five articles suffers from reference rot. *PloS one* 9, 12 (2014), e115253.

[15] Steve Lawrence, Frans Coetzee, Eric Glover, Gary Flake, David Pennock, Bob Krovetz, Finn Nielsen, Andries Kruger, and Lee Giles. 2000. Persistence of information on the web: Analyzing citations contained in research articles. In *Proceedings of the ninth international conference on Information and knowledge management*. 235–242.

[16] Wayback Machine. https://web.archive.org/.

[17] Jayant Madhavan, David Ko, Łucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. 2008. Google's deep web crawl. *VLDB* 1, 2 (2008), 1241–1252.

[18] John Markwell and David W Brooks. 2003. "Link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education* 31, 1 (2003), 69–72.

[19] Catherine C Marshall and Frank M Shipman. 2012. On the institutional archiving of social media. In *ACM/IEEE Joint Conference on Digital Libraries*.

[20] Allison McDonald, Matthew Bernhard, Luke Valenta, Benjamin VanderSloot, Will Scott, Nick Sullivan, J Alex Halderman, and Roya Ensafi. 2018. 403 forbidden: A global view of CDN geoblocking. In *IMC*.

[21] nexB/python-publicsuffix2: A small Python library to deal with publicsuffix data (includes a bundled PSL as "package data") in a wheel friendly format. Fork and continuation of Tomaž Šolc's "publicsuffix". https://github.com/nexb/python-publicsuffix2.

[22] Sandeep Pandey and Christopher Olston. 2008. Crawl ordering by search impact. In *WSDM*.

[23] Ailsa Parker. 2007. Link rot: How the inaccessibility of electronic citations affects the quality of New Zealand scholarly literature. *New Zealand Library & Information Management Journal* 50, 2 (2007), 172–192.

[24] Uri Schonfeld and Narayanan Shivakumar. 2009. Sitemaps: Above and beyond the crawl of duty. In *WWW*.

[25] Diomidis Spinellis. 2003. The decay and failures of web references. *Commun. ACM* 46, 1 (2003), 71–77.

[26] Thomas Vissers, Wouter Joosen, and Nick Nikiforakis. 2015. Parking sensors: Analyzing and detecting parked domains. In *NDSS*.

[27] InternetArchiveBot. https://meta.wikimedia.org/wiki/InternetArchiveBot.

[28] InternetArchiveBot/How the bot fixes broken links. https://meta.wikimedia.org/wiki/InternetArchiveBot/How_the_bot_fixes_broken_links.

[29] 1983–84 French Rugby Union Championship - Wikipedia. https://en.wikipedia.org/wiki/1983%E2%80%9384_French_Rugby_Union_Championship.

[30] 39 Stripes. https://en.wikipedia.org/w/index.php?title=39_Stripes&oldid=861122903.

[31] Category:Articles with permanently dead external links - Wikipedia. https://en.wikipedia.org/wiki/Category:Articles_with_permanently_dead_external_links.

[32] Size of Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

[33] Jonathan Zittrain, Kendra Albert, and Lawrence Lessig. 2014. Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Legal Information Management* 14, 2 (2014), 88–99.