# Lost, but Preserved – A Web Archiving Perspective on the Ephemeral Web

Sawood Alam
Internet Archive
San Francisco, CA, USA
sawood@archive.org

Mark Graham
Internet Archive
San Francisco, CA, USA
mark@archive.org

## ABSTRACT

The World Wide Web, our era's most dynamic information ecosystem, is characterized by its transient nature. Recent studies have highlighted the alarming rate at which web content disappears or changes, a phenomenon known as "link-rot". A 2024 Pew Research Center study revealed that 38% of webpages from 2013 were inaccessible a decade later and a quarter of the URLs from their entire dataset spanning across a decade were found dead. Even more striking, Ahrefs, an SEO company, reported that at least 66.5% of links to sites created in the last nine years are now dead. These findings echo earlier research by Zittrain et al., which uncovered significant link-rot in journalistic references from New York Times articles.

While these statistics paint a grim picture of digital impermanence, they often overlook a crucial factor: the role of web archives. This work aims to reframe the link-rot discussion by considering the preservation efforts of various web archiving institutions. Our research revisiting the Pew dataset yielded a surprising discovery: only one in ten URLs from the original study were truly missing as opposed to one in four, the remaining bulk had at least one capture in the Wayback Machine. This finding suggests that the digital landscape, when viewed through the lens of web archiving, may be less ephemeral than commonly perceived.

## KEYWORDS

Link-Rot, Web Archiving, Ephemeral Web, PEW Research, Wayback Machine

## 1 INTRODUCTION

Last year, the Pew Research Center published a link-rot study, "When Online Content Disappears" [1]. They stated, "38% of webpages that existed in 2013 are no longer accessible a decade later". They further noted, "a quarter of all webpages that existed at one point between 2013 and 2023 are no longer accessible". This is not an isolated report that quantified the rate of loss of the online information. Numerous other link-rot studies in the last two decades have reported similar numbers or worse, depending on the context and samples. For example, Ahrefs, an SEO company, earlier this year reported, "At Least 66.5% of Links to Sites in the Last 9 Years Are Dead" [2]. In 2021, Jonathan Zittrain published an article in the Atlantic, "The Internet Is Rotting" [3], in which their team analyzed about 2 million external links from the New York Times (NYTimes)[1] articles and reported that 25% of deep links have rotted. They further noted that 72% of the older links from 1998 were dead. A recent longitudinal study on link-rot from the Old Dominion University (ODU), "Some URLs Are Immortal, Most Are Ephemeral" [4], analyzed 27.3 million URL samples from the Wayback Machine[2] since 1996 and reported that about 65% of the sampled URLs were found dead on the live web, when checked in 2023. Brewster Kahle, the founder of the Internet Archive, has been citing numbers from the early days of the web and stating the average life of web pages to be anywhere from 40 to 100 days. Different studies have looked at the problem from different perspectives and contexts, hence it is often difficult to compare them side-by-side, but they all agree on the fact that an increasing number of links are rotting with the passage of time. However, some of these studies (not all) have failed to acknowledge the existence of web archives, such as the Wayback Machine, where a portion of the dead web might be preserved and can be used as a fallback when a reference leads to a broken link.

In this work we go through some of the link-rot studies and look at them from the perspective of the Wayback Machine to see how much of the dead web can be rescued. Table 1 shows the status of the dead and rescued web at a glance as sampled by a few different studies. This work was presented at the IIPC WAC 2025 conference[3] and the recording of the talk is available on YouTube[4].

## 2 METHODOLOGY

Below are brief descriptions of some terminologies that we use in this work:

- *Alive*: URLs that return `200 OK` HTTP status code when resolved
- *Dead*: URLs that return an HTTP error status codes, TCP connection errors, or DNS failures when resolved
- *Preserved*: URLs that are *Alive* on the live web as well as present in a web archive
- *Rescued*: URLs that are *Dead* on the live web, but are present in a web archive
- *Endangered*: URLs that are *Alive* on the live web, but are not present in any web archive
- *Vanished*: URLs that are *Dead* on the live web and also not present in any web archive
- *Archived*: *Preserved* + *Rescued*
- *Accessible*: *Preserved* + *Rescued* + *Endangered*

### 2.1 Datasets

Pew Research Center has generously shared their dataset with us. Their dataset contains 5.4 million unique URLs in general, news, government, and Wikipedia references categories sampled from the CommonCrawl archive[5] and Wikipedia pages. They also reported

[1] https://www.nytimes.com/

[2] https://web.archive.org/

[3] https://netpreserve.org/ga2025/

[4] https://www.youtube.com/watch?v=gmU3vFbs2GM

[5] https://commoncrawl.org/

**Table 1: Dead links from various link-rot studies rescued by the Wayback Machine.**

| Study | Year | Sample Period | Sample Size (URLs) | Dead | Rescued |
|---|---|---|---|---|---|
| **Pew (All)** | 2024 | 2013-2023 | 5.4M | 26% | 16% |
| **Pew (General)** | 2024 | 2013-2023 | 1M | 27% | 13% |
| **Zittrain NYT** | 2021 | 2013-2013 | 88K | 40% | 38% |
| **ODU NYPW** | 2024 | 1996-2021 | 27.3M | 65% | 65% |

on Tweets in their post, but that dataset was not shared with us due to the restrictions posed by the usage policies. Each URL had its categories, live status in the form of HTTP status code (including TCP or DNS failures), terminal URL and status code in case of redirects, and the year it was sampled from. The original dataset was stored in Parquet files, so performed some extractions and transformations to to it to suit our process. Then we checked URLs against the Wayback Machine to see if and when each of those were archived the first time and recorded this information in the dataset. In our study we analyzed this dataset in two forms: 1) *All* 5.4 million URLs together and 2) one million of *General* sample only.

We requested access to the dataset of about 2 million URLs from the Zittrain's NYTimes outlinks study, but could not get it. Hence, we created our own dataset by downloading all the NYTimes pages published in 2013 that are present in the Wayback Machine, extracting all the outlinks from them, and excluding all the links to pages from NYTimes itself. We were able to collect about 88 thousand such URLs this way. Then we checked the live web status of each of the URLs (after following up to 5 redirects, if any). Then we checked for their presence in the Wayback Machine. One noticeable difference from the original dataset in that we only sampled data from one year while the original dataset included URLs from a span of a decade.

We reported findings of ODU's "Not Your Parents' Web" research here directly, without any further analysis, as the work already covered what would be useful here and we were collaborators in that work.

## 3 EVALUATION

Below, we look at four different samples/studies to see how much of those samples are present in the Wayback Machine.

### 3.1 Pew (All)

When we do not take any web archives into account, about a quarter of all the 5.4 million sampled URLs would be considered inaccessible or *Dead* as illustrated in Figure 1. However, **when we leverage the Wayback Machine to access otherwise *dead* URLs, the fraction of inaccessible or *vanished* URLs drops from one in every four down to only one in every ten.** The Wayback Machine has about 72% of the entire dataset *archived*, of which 56% are *preserved* from the URLs that are still *alive* on the live web and 16% are *rescued* from the *dead*. There are 18% of the URLs from the sample that are still *alive*, but have not been *archived* in the Wayback Machine yet, which we call *endangered*, as they may become *vanished* if they cease to exist on the live web ever. It is worth noting that we did not account for any captures of these URLs that might be present in any of the many smaller web archives beyond the Wayback Machine, which if accounted for, might increase the percentage of the *accessible* URLs a little more. Moreover, we relied on HTTP status codes and did not look into the contents of the pages to check for any *soft-404s* [5] or other irrelevant content, which might change the numbers further.

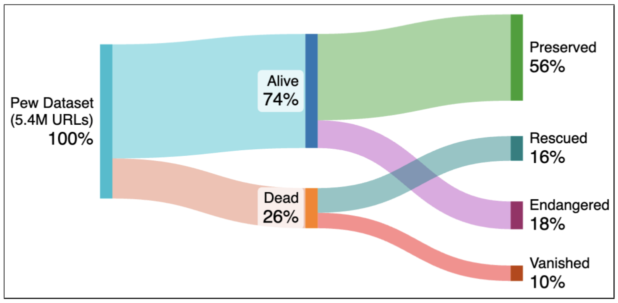

**Figure 1: Archiving status of all the URLs from the Pew dataset in the Wayback Machine.**

### 3.2 Pew (General)

A subset of about 1 million URLs from the Pew dataset is a sample of general web pages from the last decade, spanning across 11 years from 2013 to 2023. They noted that about a quarter of the URLs from this subset were *dead* in 2023, with older URLs having a greater percentage of loss, all the way to 38% for links from 2013. We recreated their yearly graph in Figure 2 in orange color with an overlay of *rescued* URLs by the Wayback Machine in green color. We found that **about 38% of the 38% *dead* URLs from 2013 (i.e., about 15% of the total) are *rescued* by the Wayback Machine.** Moreover, about a quarter of the accumulative URLs of the general sample which were considered *dead*, about half of them were *rescued* by the Wayback Machine. It is worth noting that the last three years in Figure 2 seem to be rescued almost completely, but it is due to some data contamination as we have started ingesting CommonCrawl data from the recent years into the Wayback Machine, which happens to be the source of the sample of the Pew dataset.

### 3.3 Zittrain NYT

We then looked at the dataset we created from the archived pages of NYTimes. We found that **40% of the external links from NYTimes pages from 2013 were found *dead* on the live web, but 96% of the URLs are *archived* in the Wayback Machine.** This means, only about 2% URLs from this sample have *vanished*. However, this impressive number needs to be taken with a grain of salt because we do not have the original URL sample and our

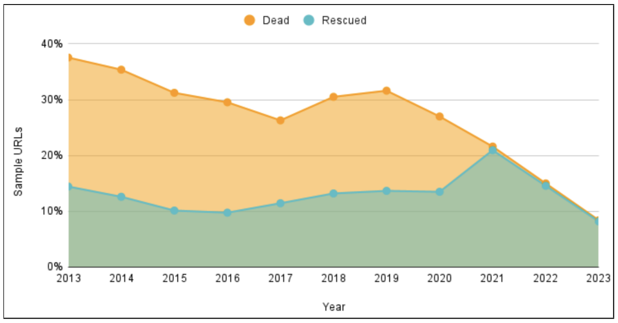

**Figure 2: Yearly archiving status of URLs from the general sample of the Pew dataset in the Wayback Machine.**

own sample is derived from pages present in the Wayback Machine, which has an inherent bias of outlinks from those pages being more likely to be archived than the outlinks of the pages that are not present in the Wayback Machine. That said, we will be keen to revisit these numbers if and when we get access to the original sample of URLs used in Zittrain's study.

## 3.4 ODU NYPW

A recent, and perhaps the most comprehensive, longitudinal link-rot study from ODU, to which we are a collaborator, analyzed 27.3 million URLs sampled from the index of the Wayback Machine spanning over more than two and a half decades. They reported **about 65% of the sampled URLs from 1996 to 2021 were found *dead* in 2023.** A significant number of these samples were not even resolving the DNS, indicating that many of those domain names were not registered anymore. They found that most of the URLs die rapidly in the first few years of their existence, but some of the longest living sites are not *dead* yet. Luckily, **all of the *dead* URLs in this sample are *rescued* by the Wayback Machine** by the virtue of it being the source of the sample in the first place. This also means, the ODU study would not be able to tell the percentage of *vanished* URLs.

## 4 CONCLUSIONS AND FUTURE WORK

In summary, all of the link-rot studies, with varying numbers, indicate that the web is brittle and an increasing number of web resources die with the passage of time. However, we found that web archives like **the Wayback Machine play an increasingly important role in *rescuing* the *dead* web and minimizing the fracture of the knowledge graph on the web**, but there is a lot more to do. For example, **the Turn All References Blue (TARB) project has fixed more than 23 million broken links (and counting) on hundreds of wikis** with the help of the InternetArchiveBot[6] and the Wayback Machine.

While there is not a lot that can be done to resurrect the *vanished* web other than attempting to find alternate locations where the content might have moved to (via projects like FABLE[7]), we are determined to minimize the percentage of the endangered URLs. However, there are some internal and external factors that limit our ability to make it ZERO, such as, resource limitations, JavaScript-heavy pages, bot blocking, loginwalls, paywalls, deepweb, lack of timely discovery, etc. We strive to narrow down the potential loss of our cultural heritage via different means such as ingesting feeds from MediaCloud[8], GDELT[9], Wikipedia EventStream[10], and more recently, becoming part of the IndexNow[11] initiative for link discovery soon after corresponding page creation or update on the web.

## 5 ACKNOWLEDGMENTS

We thank our friends at the Pew Research Center and the Old Dominion University, our colleagues Jake LaFountain and Stephen Balbach, and our intern Rachel Auslander for their help and support in this work.

## REFERENCES

[1] A. Chapekis, S. Bestvater, E. Remy, and G. Rivero, "When Online Content Disappears," https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/, 2024.
[2] P. Stox, M. Pecánek, and J. Hardwick, "At Least 66.5% of Links to Sites in the Last 9 Years Are Dead," https://ahrefs.com/blog/link-rot-study/, 2024.
[3] J. L. Zittrain, "The Internet Is Rotting," https://www.theatlantic.com/technology/archive/2021/06/the-internet-is-a-collective-hallucination/619320/, 2024.
[4] M. C. Weigle, K. Garg, S. Alam, D. Ayala, and M. L. Nelson, "Some URLs Are Immortal, Most Are Ephemeral," https://ws-dl.blogspot.com/2024/09/2024-09-20-some-urls-are-immortal-most.html, 2024.
[5] L. Meneses, R. Furuta, and F. Shipman, "Identifying 'Soft 404' Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections," in *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, ser. TPDL '12, vol. 7489, 2012, pp. 197–208.

[6] https://meta.wikimedia.org/wiki/InternetArchiveBot
[7] https://webresearch.eecs.umich.edu/fable/
[8] https://www.mediacloud.org/
[9] https://www.gdeltproject.org/
[10] https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams_HTTP_Service
[11] https://www.indexnow.org/