

Lost, But Preserved

A Web Archiving Perspective on the Ephemeral Web

Sawood Alam & Mark Graham
Wayback Machine, Internet Archive

[@ibnesayeed](https://twitter.com/ibnesayeed) [@waybackmachine](https://twitter.com/waybackmachine) [@internetarchive](https://twitter.com/internetarchive)
@sawood@mastodon.archive.org



“38% of webpages that existed in 2013 are no longer accessible a decade later”

– Pew Research

Link Rot and Digital Decay on

pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/ Incognito (2) New Chrome available

NUMBERS, FACTS AND TRENDS SHAPING YOUR WORLD NEWSLETTERS | PRESS | MY ACCOUNT | DONATE | CONTACTED BY US?

Read our research on: Congress | Donald Trump | Religion

Pew Research Center

Search pewresearch.org... 

RESEARCH TOPICS ▾ PUBLICATIONS OUR METHODS SHORT READS TOOLS & RESOURCES EXPERTS ABOUT US

Home > Research Topics > Internet & Technology

REPORT | MAY 17, 2024 SHARE 

When Online Content Disappears

38% of webpages that existed in 2013 are no longer accessible a decade later

BY ATHENA CHAPEKIS, SAMUEL BESTVATER, EMMA REMY AND GONZALO RIVERO

How we did this 

The internet is an unimaginably vast repository of modern life, with hundreds of billions of indexed webpages. But even as users across the world rely on the web to access books, images, news articles and other resources, this content sometimes disappears from view.

REPORT MATERIALS

Report PDF

TABLE OF CONTENTS

When Online Content Disappears

- Webpages from the last decade
- Links on government websites
- Links on news websites
- Reference links on Wikipedia
- Posts on Twitter

<https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/>

**“At least
66.5% of
links to sites
in the last 9
years are
dead”**

— Ahrefs

The screenshot shows a web browser window with the URL ahrefs.com/blog/link-rot-study/. The page has a blue header with the Ahrefs logo and navigation links for SEO, Marketing, Data & Studies, Product, Guides, and a search bar. Below the header is a cartoon illustration of a man with orange hair and glasses, wearing a green top hat and a black t-shirt, juggling three rings. The main title of the article is "At Least 66.5% of Links to Sites in the Last 9 Years Are Dead (Ahrefs Study on Link Rot)". The author is listed as Patrick Stox, and it's reviewed by Michal Pecánek and Joshua Hardwick. The date is February 2, 2024, and the read time is 8 minutes.

<https://ahrefs.com/blog/link-rot-study/>

“25% of about
2 million
external deep
links from
NYTimes
articles have
rotted”

— Jonathan Zittrain

The screenshot shows a web browser window for The Atlantic. The URL in the address bar is theatlantic.com/technology/archive/2021/06/the-internet-is-a-collective-hallucination/619320/. The page title is "The Internet Is Rotting" by Jonathan L. Zittrain. The subheading reads: "Too much has been lost already. The glue that holds humanity's knowledge together is coming undone." Below the author's name is a large, mostly blank image area. The top navigation bar includes links for Popular, Latest, Newsletters, and a sign-in/subscribe button.

<https://www.theatlantic.com/technology/archive/2021/06/the-internet-is-a-collective-hallucination/619320/>

“A year after
the Egyptian
Revolution,
10% of the
social media
documentation
is gone”

— Hany SalahEldeen

2012-02-11: Losing My Revolution x +

ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html Incognito (2) New Chrome available :

← Web Science and Digital Libraries Research Group

Research and Teaching Updates from the Web Science and Digital Libraries Research Group (@WebSciDL) at Old Dominion University.

2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.

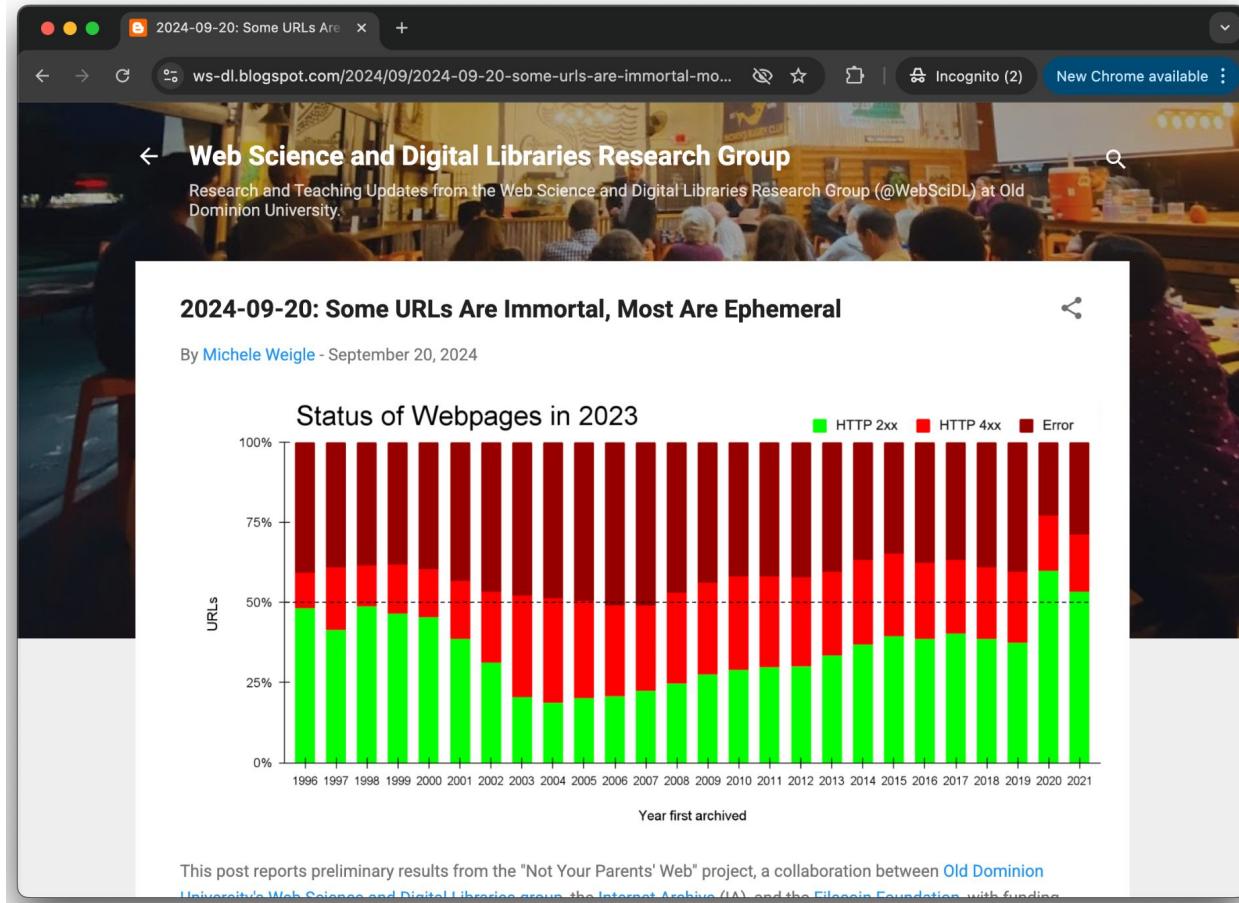
By Hany SalahEldeen - February 11, 2012

The Egyptian revolution on the 25th of January 2011 was unlike any other revolution in history because of the role of social media. Several blogs, Storify entries, web pages, channels on YouTube where created to document the revolution. Several books were even published documenting the 18 days. All of these contributions were made by the public, not historians, utilizing the tools of web 2.0. As a result of all these contributions we have an enormous digital content

<https://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html>

“Only 35% of webpages sampled from a span of 25 years were still alive in 2023”

– WSDL ODU



“One in five scholarly articles suffers from reference rot”

– Klein et al.

The screenshot shows a PLOS One article page. The title is "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot" by Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. It was published on December 26, 2014. The page includes sections for Article, Authors, Metrics, Comments, and Media Coverage. The Abstract section discusses the impact of reference rot on scholarly communication. The URL of the article is <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253>.

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253>

The average lifetime of a webpage is 40, 75, or 100 days – Brewster Kahle

1996

Scientific American: Article... | Brewster Kahle . In Scientific American

web.archive.org/web/19970215093036/http://www.sciam.com/0397issue/0397kahle.html

155 captures | 15 Feb 1997 - 28 Feb 1997

SCIENTIFIC AMERICAN

CURRENT ISSUE | INTERVIEW | MARKETPLACE | ASK THE EXPERTS | FEEDBACK | BOOKMARKS | SEARCH | MAIN MENU

SPECIAL REPORT

Preserving the Internet

An archive of the Internet may prove to be a vital record for historians, businesses and governments

by [Brewster Kahle](#)

SUBTOPICS: [Keeping Missing Links](#)

Manuscripts from the library of Alexandria in ancient Egypt disappeared in a fire. The early printed books decayed into unrecognizable shreds. Many of the oldest cinematic films were recycled for their silver content. Unfortunately, history may repeat itself in the evolution of the Internet--and its World Wide Web.

No one has tried to capture a comprehensive

1996

Brewster Kahle . In Scientific American | On the Web, Research Work Proves Ephemeral

http://19971011050140/http://www.archive.org/sciam_article.html

Archiving the Internet

SCIENTIFIC AMERICAN

Brewster Kahle
Internet Archive
11/4/96

Bold efforts to record the entire Internet are expected to lead to new services.

Submitted to Scientific American for March 1997 Issue

2003

washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral-959c882f-9ad0-4b36-88cd-fb7411db118d/

The Washington Post

Democracy Dies in Darkness

This article was published more than 21 years ago

Politics | Donald Trump | The Fix | The Briefs | Polling | Democracy in America | Elections

On the Web, Research Work Proves Ephemerical

Electronic Archivists Are Playing Catch-Up in Trying to Keep Documents From Landing in History's Dustbin

November 24, 2003

By Rick Weiss

It was in the mundane course of getting a scientific paper published that physician Robert Dellavalle came to the unsettling realization that the world was dissolving before his eyes.

The world, that is, of footnotes, references and Web pages.

Cut through the 2024 election noise. Get The Campaign Moment newsletter.

Dellavalle, a dermatologist with the Veterans Affairs Medical Center

Most Read Politics >

<http://www.sciam.com/0397issue/0397kahle.html>

http://www.archive.org/sciam_article.html

<https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral-959c882f-9ad0-4b36-88cd-fb7411db118d/>

Not all the web is archived equal – WSDL ODU

The screenshot shows a browser window with two tabs. The left tab is for arXiv.org with the URL arxiv.org/abs/1212.6177. The right tab is for a Cornell University page with the URL [\[1212.6177\].cornell.edu/](https://[1212.6177].cornell.edu/). The Cornell page has a red header and sidebar, while the arXiv page has a white header and sidebar.

Cornell University
We gratefully acknowledge support from the National Science Foundation.
arXiv > cs > arXiv:1212.6177
Computer Science > Digital Libraries
(Submitted on 26 Dec 2012 (v1), last revised 6 Jan 2013 (this version, v2))
How Much of the Web Is Archived?
Scott G. Ainsworth, Ahmed AlSum, Hany SalahEldeen, Michele C. Weigle, Michael L. Nelson

Although the Internet Archive's Wayback Machine is the largest and most well-known web archive, there have been a number of public web archives that have emerged in the last several years. With varying resources, audiences and collection development policies, these archives have varying levels of overlap with each other. While individual archives can be measured in terms of number of URIs, number of copies per URI, and intersection with other archives, to date there has been no answer to the question "How much of the Web is archived?" We study the question by approximating the Web using sample URIs from DMOZ, Delicious, Bitly, and search engine indexes; and, counting the number of copies of the sample URIs exist in various public web archives. Each sample set provides its own bias. The results from our sample sets indicate that range from 35%–90% of the Web has at least one archived copy, 17%–49% has between 2–5 copies, 1%–8% has 6–10 copies, and 8%–63% has more than 10 copies in public web archives. The number of URI copies varies as a function of time, but no more than 31.3% of URIs are archived more than once per month.

Comments: This is the long version of the short paper by the same title published at JCDL'11. 10 pages, 5 figures, 7 tables. Version includes minor typographical corrections
Subjects: Digital Libraries (cs.DL); Information Retrieval (cs.IR)
ACM classes: H.3.7
Cite as: [arXiv:1212.6177 \[cs.DL\]](https://arxiv.org/abs/1212.6177)
(or [arXiv:1212.6177v2 \[cs.DL\]](https://arxiv.org/abs/1212.6177v2) for this version)
<https://doi.org/10.48550/arXiv.1212.6177>

Submission history

<https://arxiv.org/abs/1212.6177>

The screenshot shows a browser window with two tabs. The left tab is for the ACM Digital Library with the URL dl.acm.org/doi/10.1145/3041656. The right tab is for a Cornell University page with the URL [\[1212.6177\].cornell.edu/](https://[1212.6177].cornell.edu/). The Cornell page has a red header and sidebar, while the ACM page has a white header and sidebar.

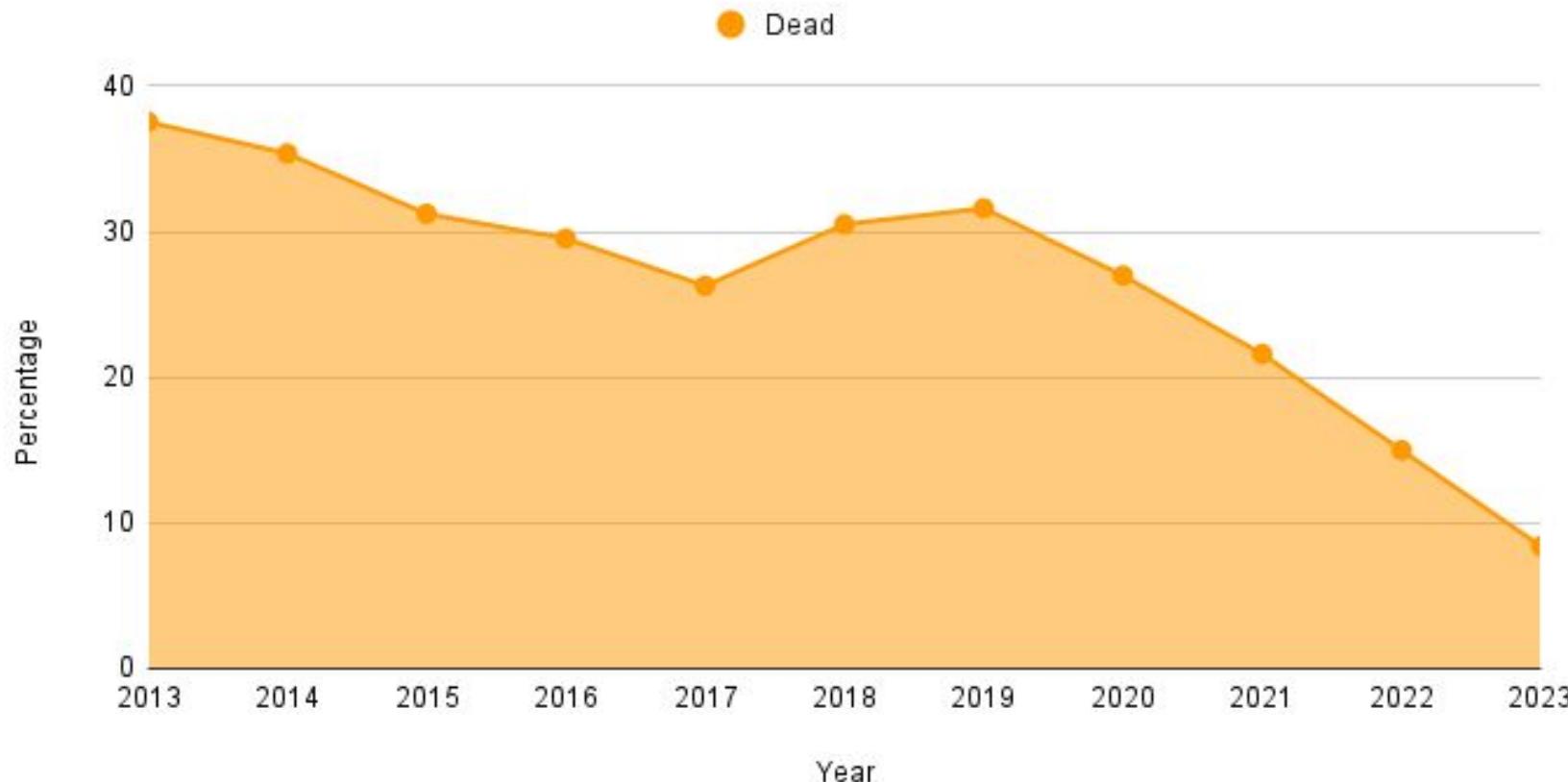
ACM DL DIGITAL LIBRARY | Association for Computing Machinery
Journals Magazines Proceedings Books SIGs Conferences People Search ACM Digital Library Advanced Search
Journal Home Just Accepted Latest Issue Archive Authors Editors Reviewers About Contact Us
Home > ACM Journals > ACM Transactions on Information Systems > Vol. 36, No. 1 > Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages
RESEARCH-ARTICLE
Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages
Authors: Lulwah M. Alkwai, Michael L. Nelson, Michele C. Weigle | [Authors Info & Claims](#)
ACM Transactions on Information Systems (TOIS), Volume 36, Issue 1 • Article No.: 1, Pages 1 - 34 • <https://doi.org/10.1145/3041656>
Published: 05 June 2017 [Publication History](#)

Abstract
It has long been suspected that web archives and search engines favor Western and English language webpages. In

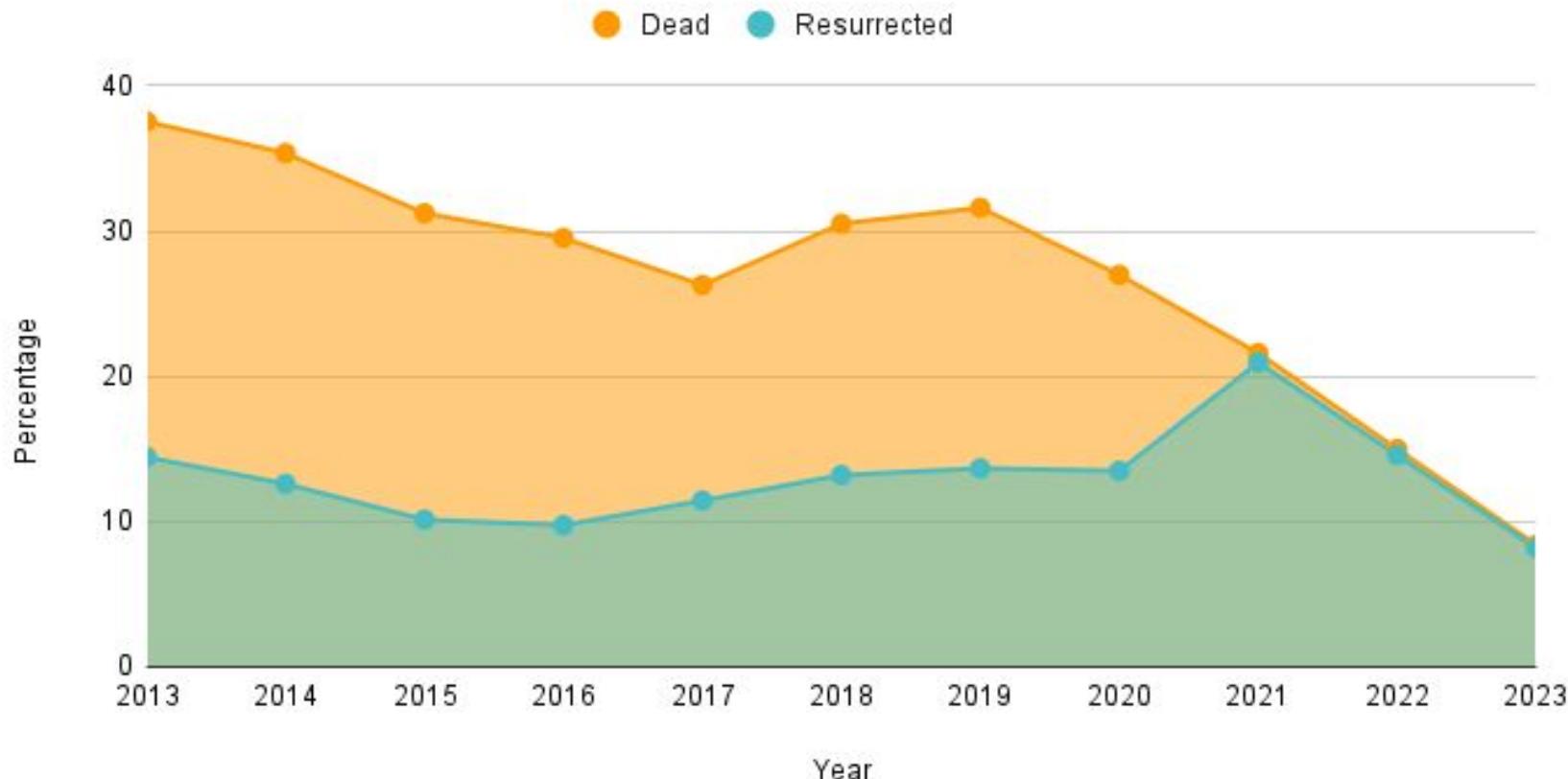
Link-rot terminologies

- **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*

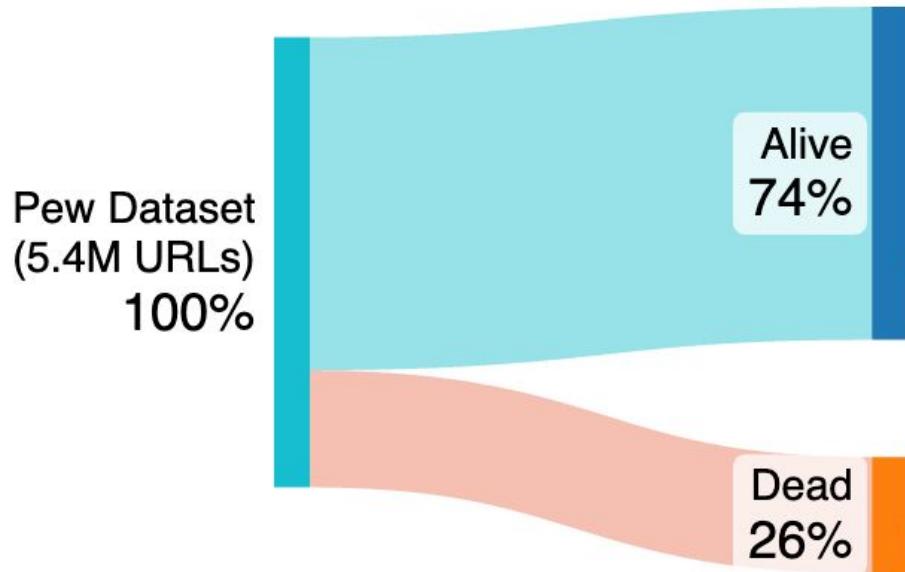
 38% of the links from 2013 are dead! – Pew Research



😊 But 38% of those are resurrected! – Wayback Machine

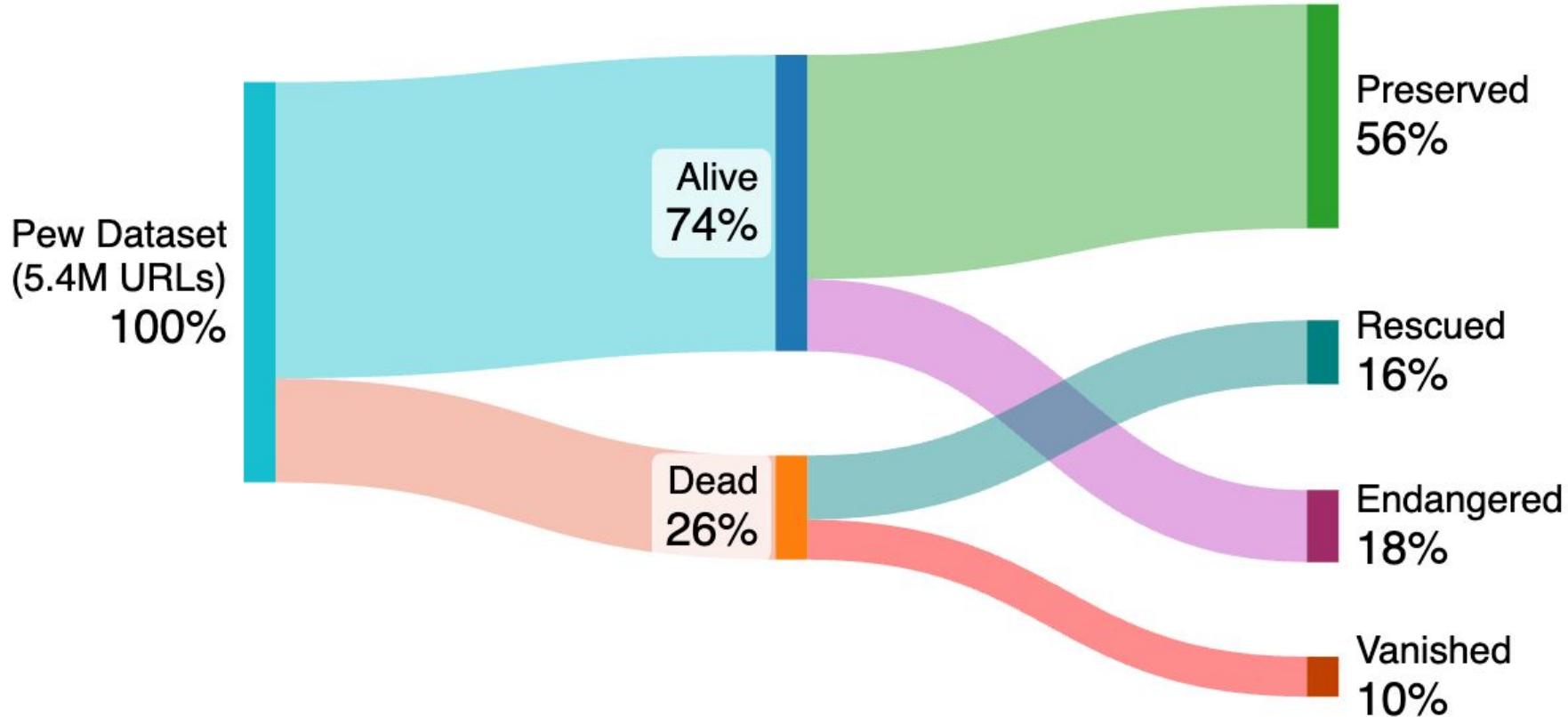


:(Every 1 in 4 URLs is dead! – Pew Research





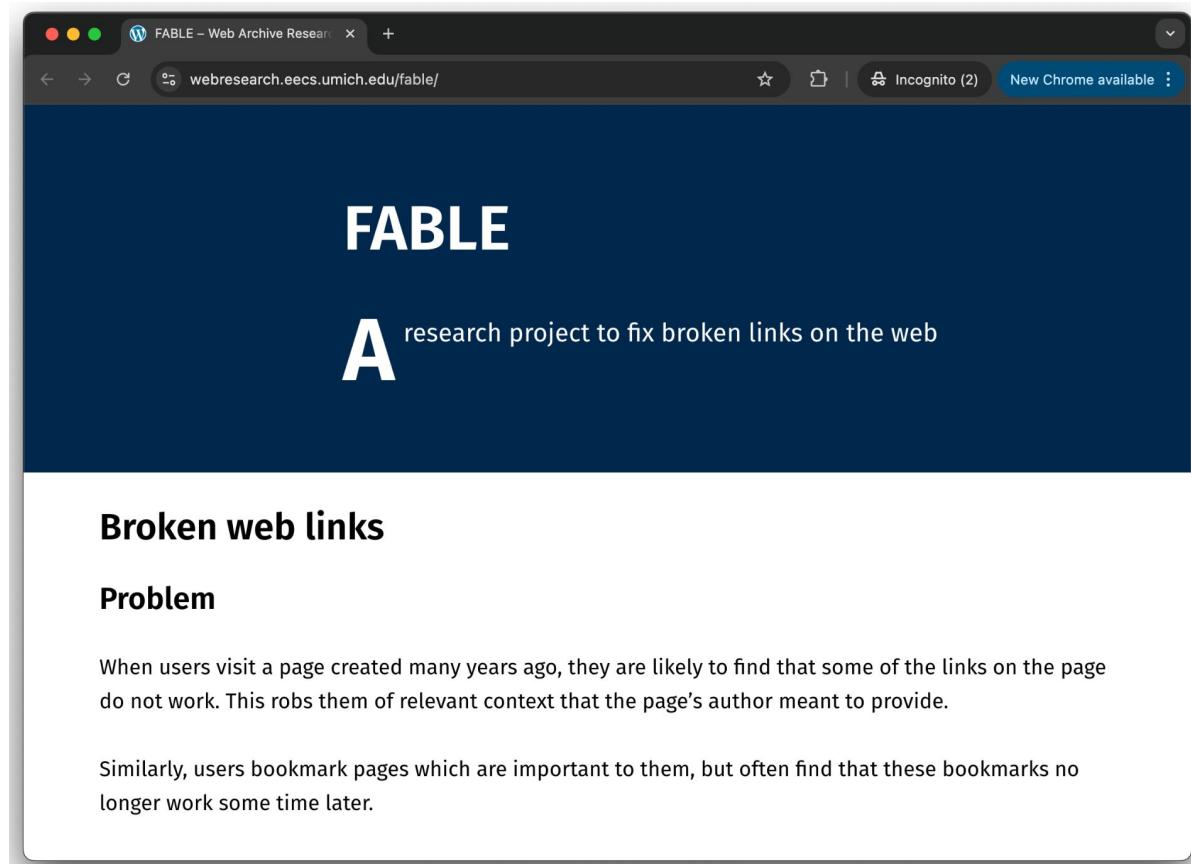
Only 1 in every 10 URLs is vanished! – Wayback Machine



Vanished web: Discover new locations for the same content

- Search engine queries
- Historical redirects
- URL change patterns

<https://webresearch.eecs.umich.edu/fable/>



The screenshot shows a web browser window with the title bar "FABLE – Web Archive Research". The address bar contains the URL "webresearch.eecs.umich.edu/fable/". The main content area has a dark blue header with the word "FABLE" in large white letters. Below the header, there is a large white letter "A" followed by the text "research project to fix broken links on the web". The main body of the page is white and features the heading "Broken web links" and a section titled "Problem". The text in the "Problem" section states: "When users visit a page created many years ago, they are likely to find that some of the links on the page do not work. This robs them of relevant context that the page's author meant to provide." Another paragraph below says: "Similarly, users bookmark pages which are important to them, but often find that these bookmarks no longer work some time later."

Vanished web: Cross-archive Memento lookup with MemGator

```
$ memgator -f cdxj http://si.edu/ | grep -v "!" | cut -d'/' -f3 | sort | uniq -c | sort -nr
13263 web.archive.org
3590 wayback.archive-it.org
1202 web.archive.bibalex.org
651 webarchive.loc.gov
321 arquivo.pt
32 wayback.vefsafn.is
11 web.archive.org.au
3 archive.is
1 www.webarchive.org.uk
1 swap.stanford.edu
1 perma.cc
$ memgator -f cdxj http://odu.edu/ | grep -v "!" | cut -d'/' -f3 | sort | uniq -c | sort -nr
3071 web.archive.org
796 wayback.archive-it.org
751 web.archive.bibalex.org
99 webarchive.loc.gov
26 arquivo.pt
2 archive.is
1 wayback.vefsafn.is
```

<https://github.com/oduwsdl/MemGator>

Challenges in preserving the endangered web

**Resource
constraints**

**JS-heavy
webpages**

**Bot
blocking**

Loginwalls

Paywalls

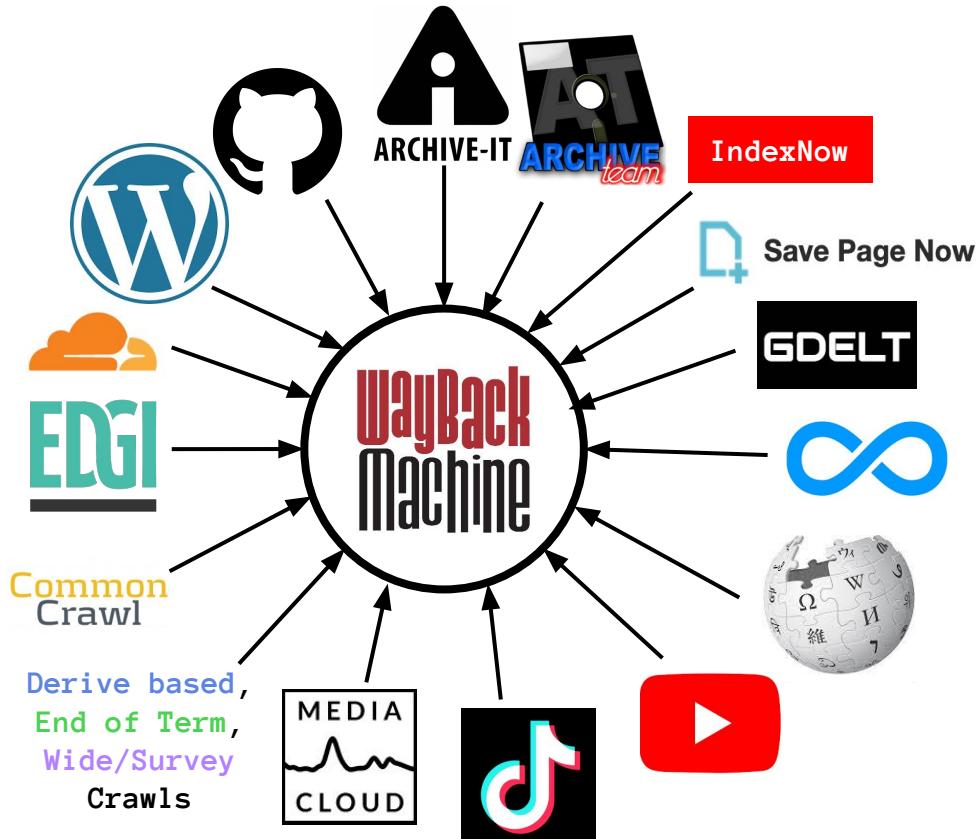
Deepweb

**Rate
limiting**

**Timely
discovery**

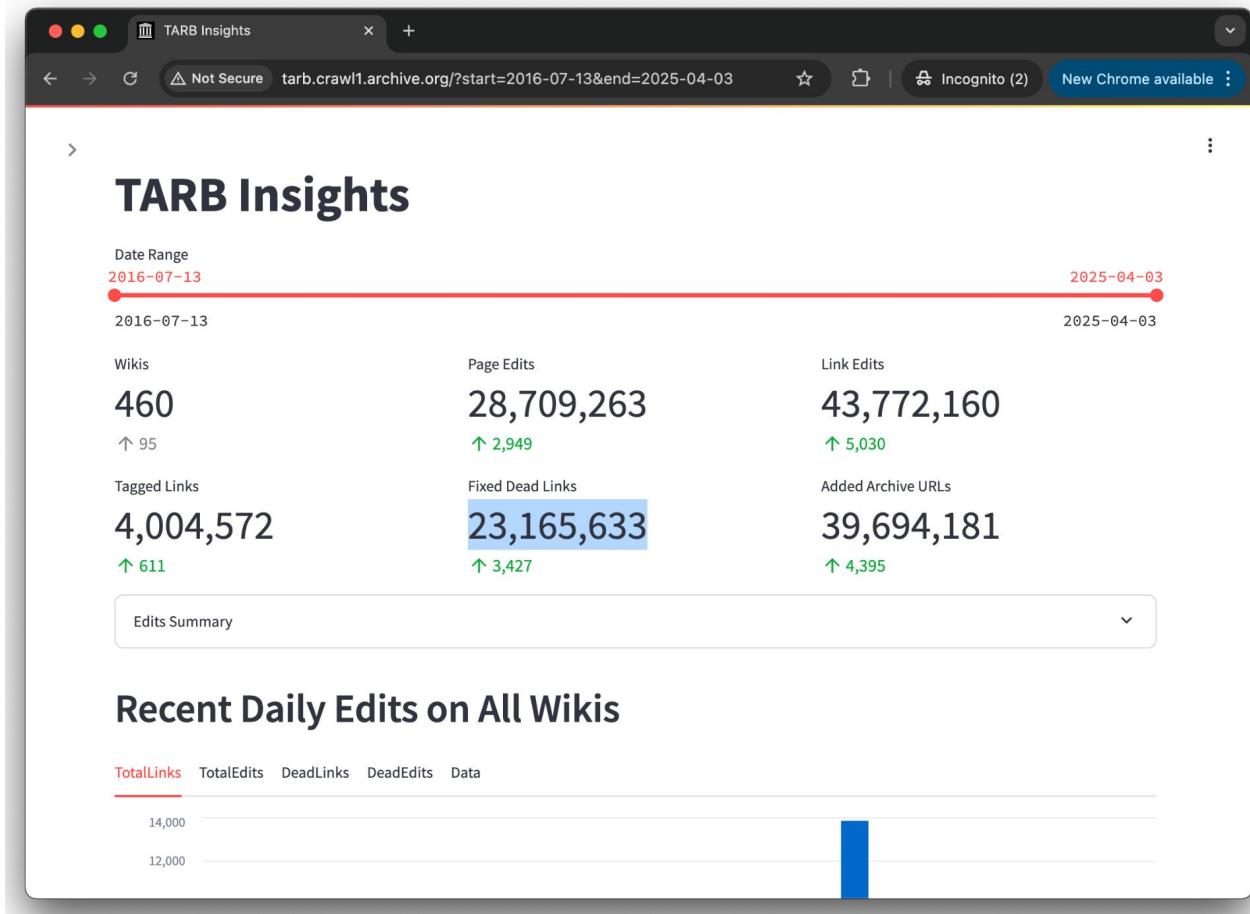
**Geo
targeting**

Endangered web: Many sources of link discovery & archiving



Over 1.8 billion URLs are added to the Wayback Machine every day

IABot has fixed over 23 million broken links on various Wikipedia language editions



Two thirds of the dead links on Wikipedia are rescued?

Unreliable!!!

The screenshot shows a web browser window titled "TARB IABot Dataset". The address bar indicates the site is "Not Secure" and the URL is "tarb.crawl1.archive.org/data". The main content area displays the title "TARB IABot Dataset" and a question: "How many total links are there, how many are dead, how many are archived, and how many are dead but also archived?". Below this is an SQL query:

```
SELECT
    COUNT(*) AS total_links,
    SUM(CASE WHEN live_state IN (0, 1, 2) THEN 1 ELSE 0 END) AS dead_links,
    SUM(CASE WHEN has_archive = 1 THEN 1 ELSE 0 END) AS archived_links,
    SUM(CASE WHEN live_state IN (0, 1, 2) AND has_archive = 1 THEN 1 ELSE 0 END) AS dead_and_archived_links;
FROM externallinks_global;
```

Below the query is a help message: "May I help? (e.g., How many dead links are there on the NASA page from the Portuguese wiki?)". A "Run Query" button is present, and the results are displayed in a table:

	total_links	dead_links	archived_links	dead_and_archived_links
0	137,256,659	19,828,483	39,359,905	13,217,217

Recreating Zittrain's NYTimes dataset

Page collection

Link extraction

Live status

Archive status

Collect 2013 NYTimes articles archived in the Wayback Machine.

Extract all the external links from those articles (88k unique URLs).

Check the live web to see how many of the external links were dead (40%).

Check the Wayback Machine to see how many are rescued (38%).

- The original, now inaccessible, dataset had about 2 million links
- We collected articles of only one year as opposed to a decade
- Our results are biased towards links of archived NYTimes pages

Acknowledgements

- ***Aaron Smith & Team, Pew Research***
 - Shared their dataset and methodology
- ***Rachel Auslander, Internet Archive Intern***
 - Catalogued some related works
- ***Jake LaFountain, Internet Archive***
 - Helped with the recreation of NYTimes dataset

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	2024	1996-2021	27.3M	65%	65%