

Lost, But Preserved

A Web Archiving Perspective on the Ephemeral Web

Sawood Alam & Mark Graham
Wayback Machine, Internet Archive

@ibnesayeed @waybackmachine @internetarchive
@sawood@mastodon.archive.org



September 15, 2025, Web Archiving and Digital Libraries (WADL) Workshop 2025, Chicago, Illinois, USA

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

– Pew Research

The screenshot shows a web browser window displaying the Pew Research Center website. The URL in the address bar is <https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/>. The page features the Pew Research Center logo and navigation links. The main content area displays the title "When Online Content Disappears" and the subtitle "38% of webpages that existed in 2013 are no longer accessible a decade later". The authors listed are Athena Chapekis, Samuel Bestvater, Emma Remy, and Gonzalo Rivero. A "How we did this" section is partially visible. On the right side, there are links to "Report Materials" (including a "Report PDF") and a "Table of Contents" listing various types of content that may be lost, such as webpages, government websites, news websites, Wikipedia links, and Twitter posts.

Link Rot and Digital Decay on x +

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

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

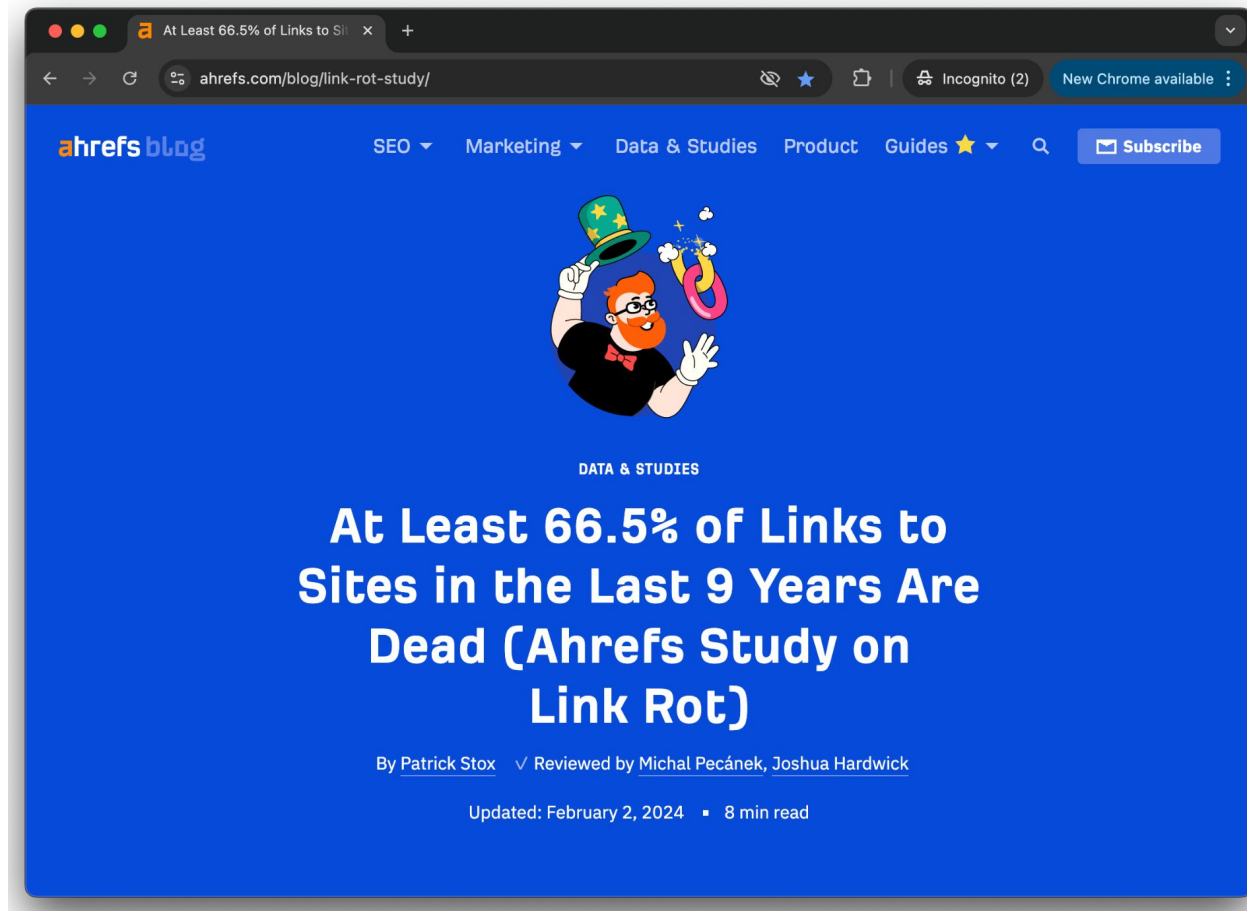
TABLE OF CONTENTS

- 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”

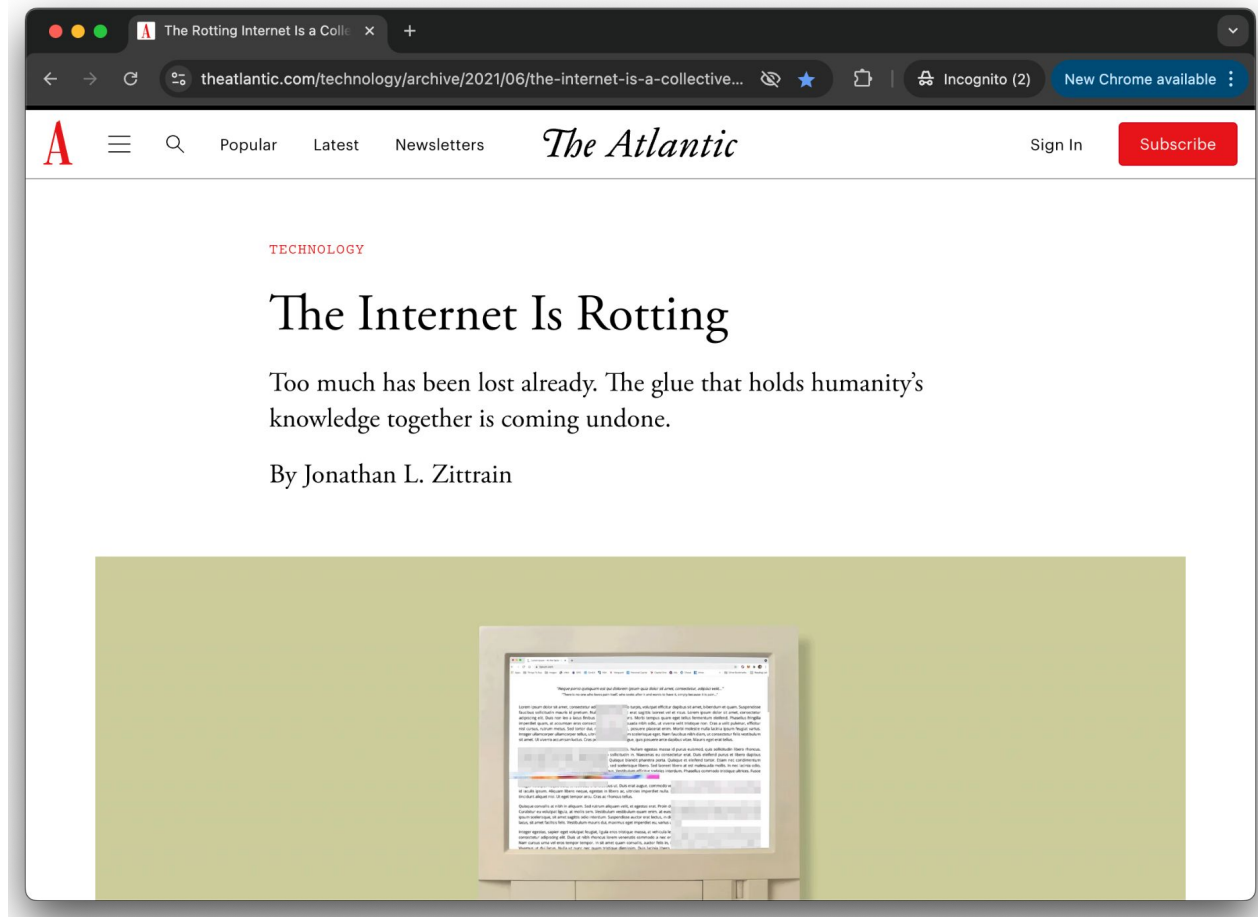
– Ahrefs



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

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

– Jonathan Zittrain



<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

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

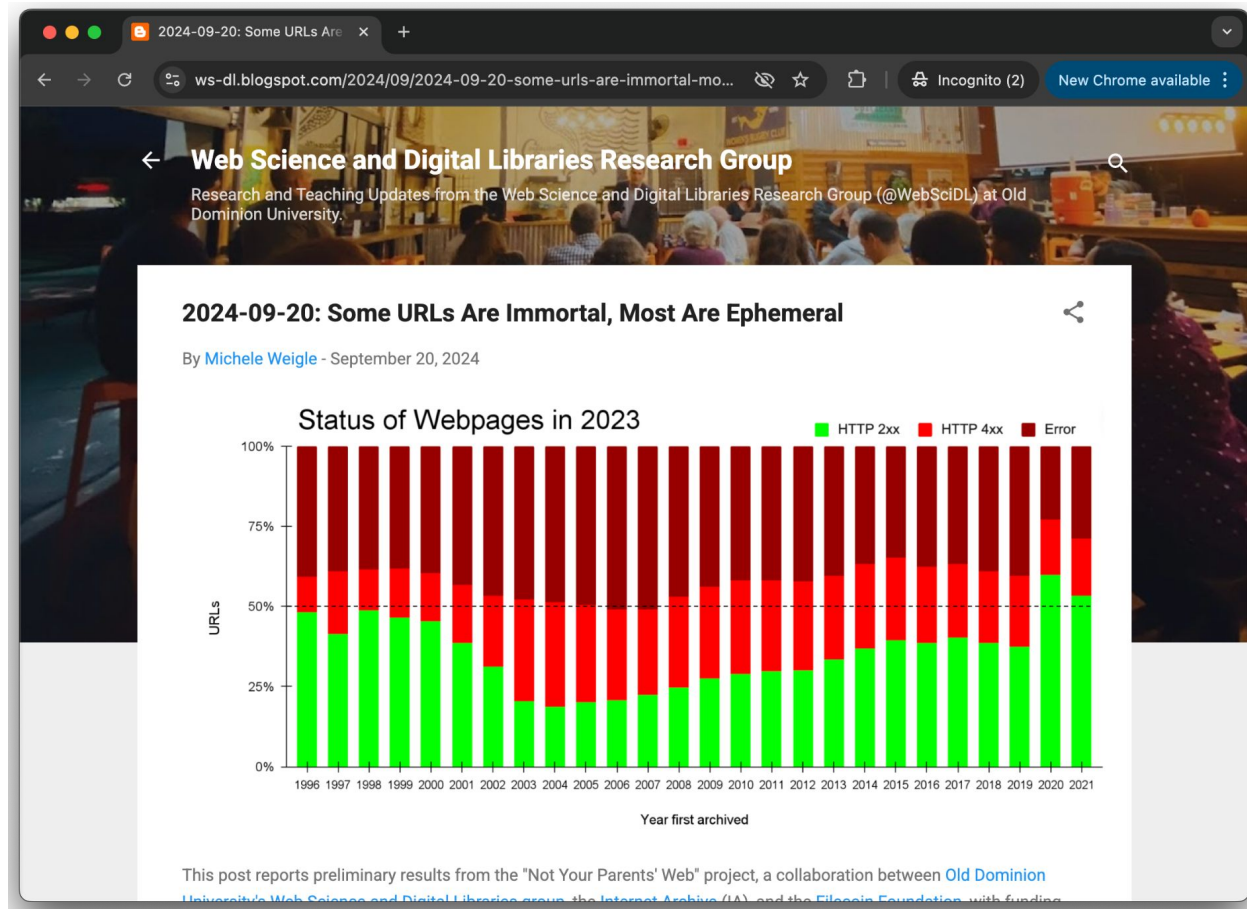


The Egyptian revolution on the 25th of January 2011 was unlike any other revolution in history because of the role of social media. Several blogs, Storify entries, web pages, channels on YouTube were created to document the revolution. Several books were even published documenting the 18 days. All of 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



<https://ws-dl.blogspot.com/2024/09/2024-09-20-some-urls-are-immortal-most.html>

“One in five
scholarly
articles
suffers from
reference
rot”

– Klein et al.

The screenshot shows a web browser window displaying a PLOS ONE article. The browser's address bar shows the URL: journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253. The page features the PLOS ONE logo, navigation links (Publish, About, Browse), and a search bar. The article 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 is an open access, peer-reviewed research article published on December 26, 2014. The article has 142 saves, 123 citations, 29,651 views, and 167 shares. A table of contents on the left lists sections: Article, Authors, Metrics, Comments, and Media Coverage. The abstract section is highlighted, showing the title "Abstract" and a brief introduction. On the right, there are buttons for "Download PDF", "Print", and "Share", along with a "Check for updates" button. A "Subject Areas" section at the bottom right lists various topics like Archives, Extrapolation, Communications, etc.

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin

Published: December 26, 2014 • <https://doi.org/10.1371/journal.pone.0115253>

Article	Authors	Metrics	Comments	Media Coverage
142 Save	123 Citation	29,651 View	167 Share	

Abstract

The emergence of the web has fundamentally affected most aspects of information communication, including scholarly communication. The immediacy that characterizes publishing information to the web, as well as accessing it, allows for a dramatic increase in the speed of dissemination of scholarly knowledge. But, the transition from a paper-based to a web-based scholarly communication system also poses challenges. In this paper, we focus on reference rot, the combination of link rot and content drift to which references to web resources included in Science, Technology, and Medicine (STM) articles are subject. We investigate the extent to which reference rot impacts the ability to revisit the web context that surrounds STM articles some time after their publication. We do so on the basis of a vast collection of articles from three corpora that span publication years 1997 to 2012. For over one million references to web resources extracted from over 3.5 million articles, we determine whether the HTTP URI is still responsive on the live web and whether web archives contain an archived snapshot.

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

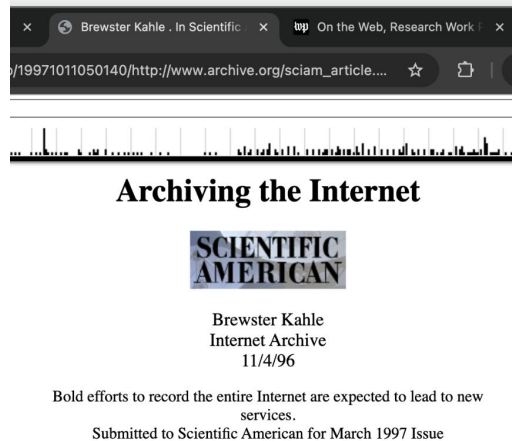
The average lifetime of a webpage is 40, 75, or 100 days – Brewster Kahle

1996



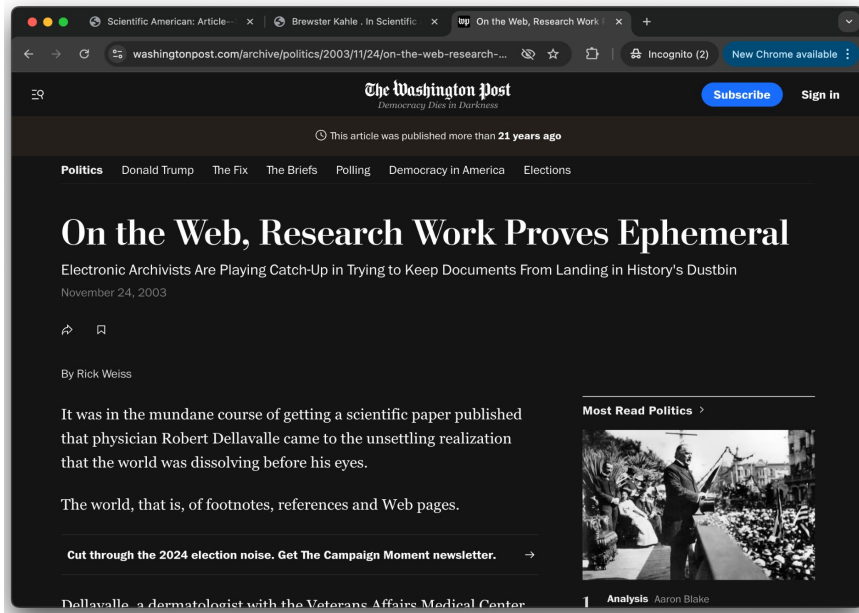
<http://www.sciam.com/0397issue/0397kahle.html>

1996



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

2003



<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

<https://dl.acm.org/doi/10.1145/3041656>

The screenshot shows the arXiv preprint page for the paper "How Much of the Web Is Archived?". The page is from the Cornell University arXiv repository, specifically in the Computer Science > Digital Libraries section. The paper was submitted on 26 Dec 2012 (v1), last revised 6 Jan 2013 (this version, v2). The authors listed are Scott G. Ainsworth, Ahmed AlSum, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. The abstract discusses the Internet Archive's Wayback Machine and compares it with other public web archives, noting varying levels of overlap and collection development policies. The comments section mentions that this is the long version of a short paper published at JCDL'11, consisting of 10 pages, 5 figures, and 7 tables. The submission history is also visible.

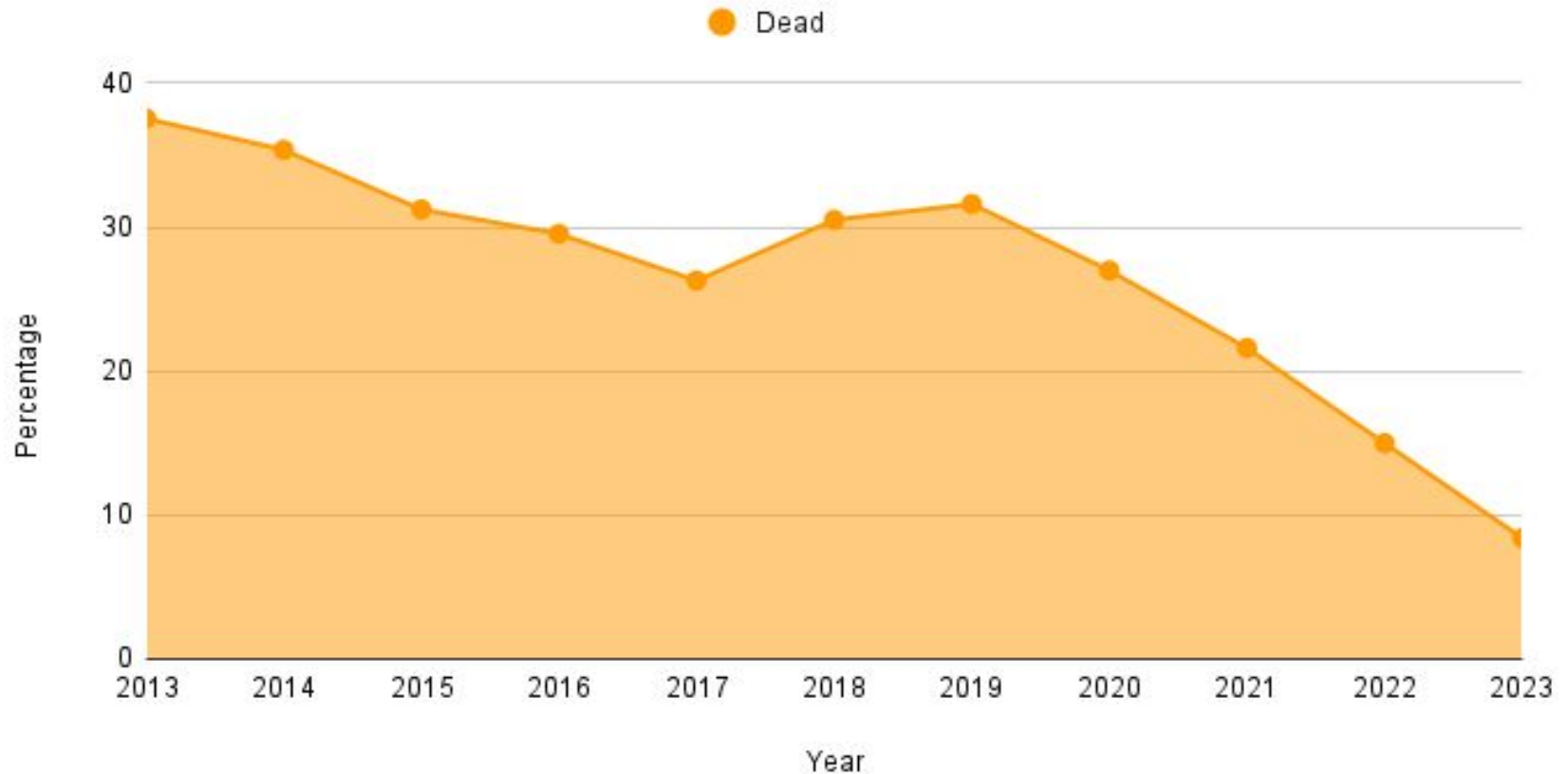
<https://arxiv.org/abs/1212.6177>

The screenshot shows the ACM Digital Library page for the paper "Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages". The paper is from the ACM Transactions on Information Systems (TOIS), Volume 36, Issue 1, Article No. 1, Pages 1 - 34. The authors listed are Lulwah M. Alkawai, Michael L. Nelson, and Michele C. Weigle. The paper was published on 05 June 2017. The abstract states: "It has long been suspected that web archives and search engines favor Western and English language webpages. In".

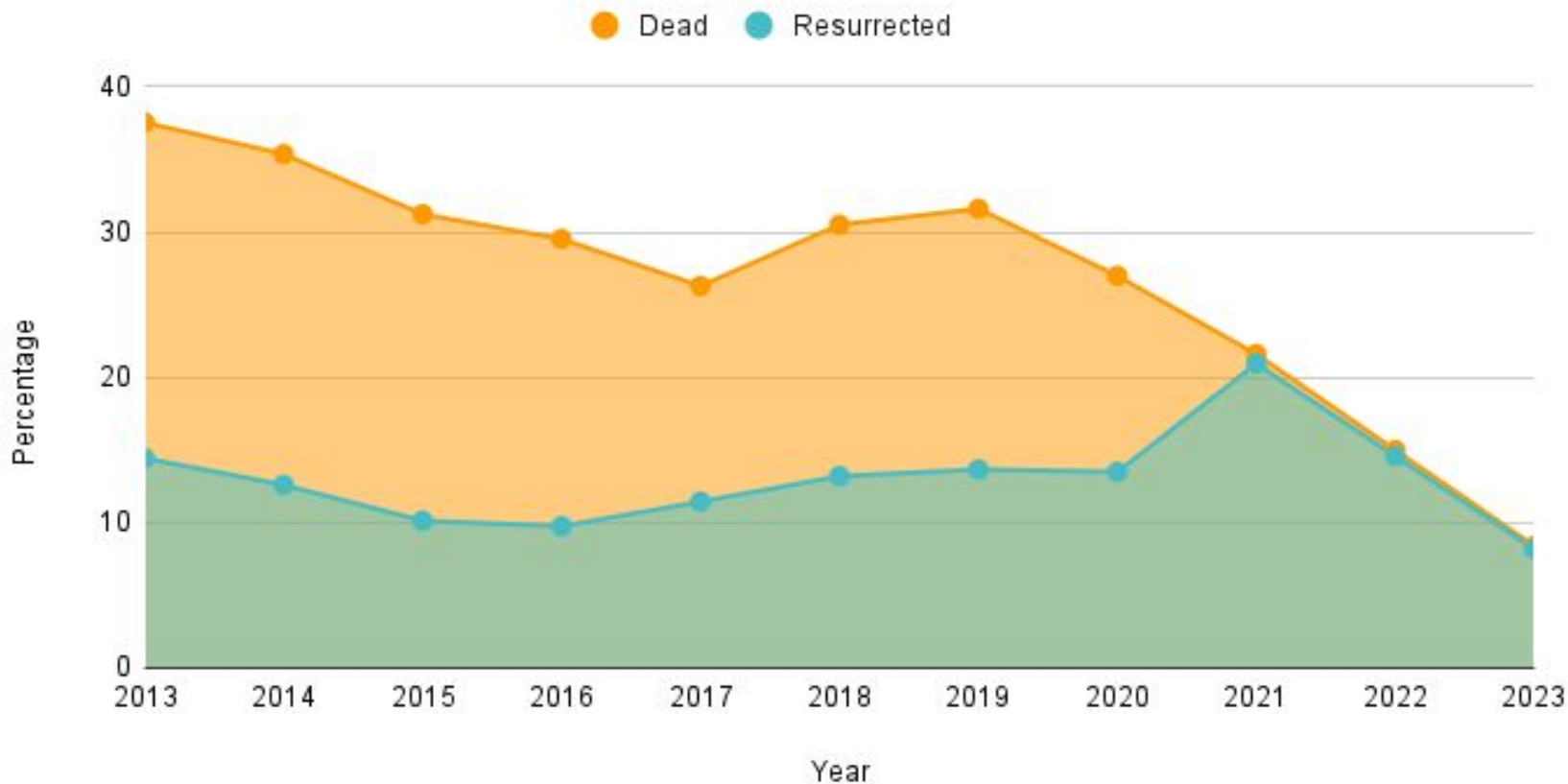
Link-rot terminologies

- **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

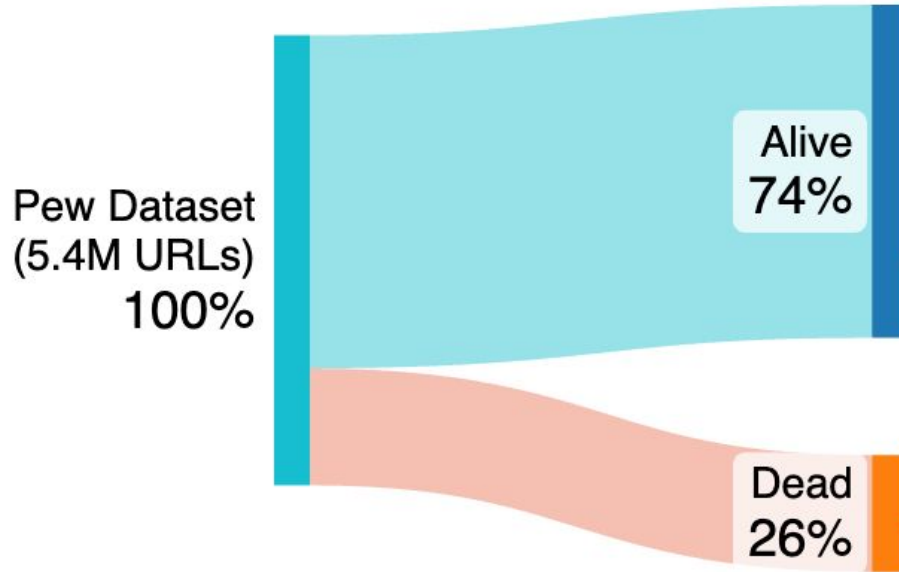


😊 But 38% of those are resurrected! – Wayback Machine

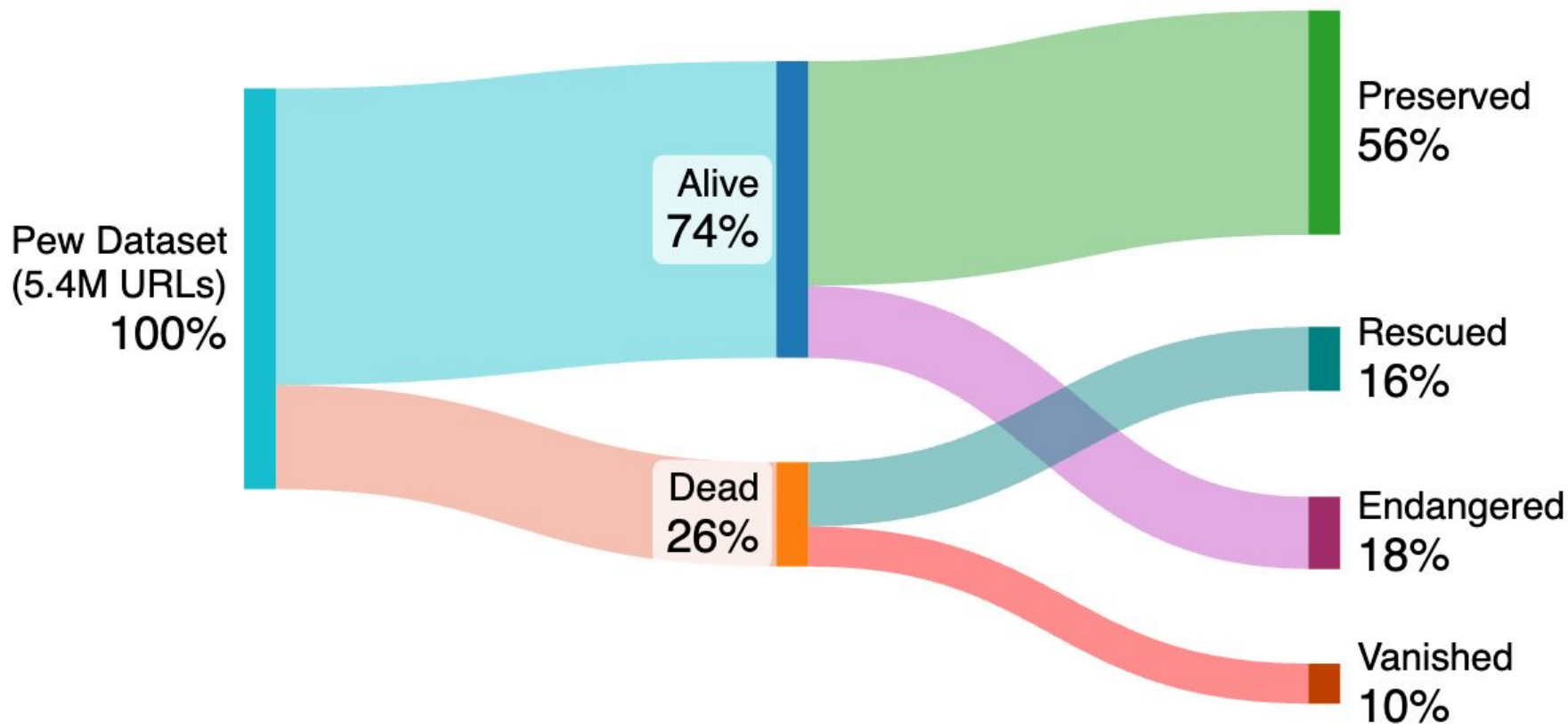




Every 1 in 4 URLs is dead! – Pew Research



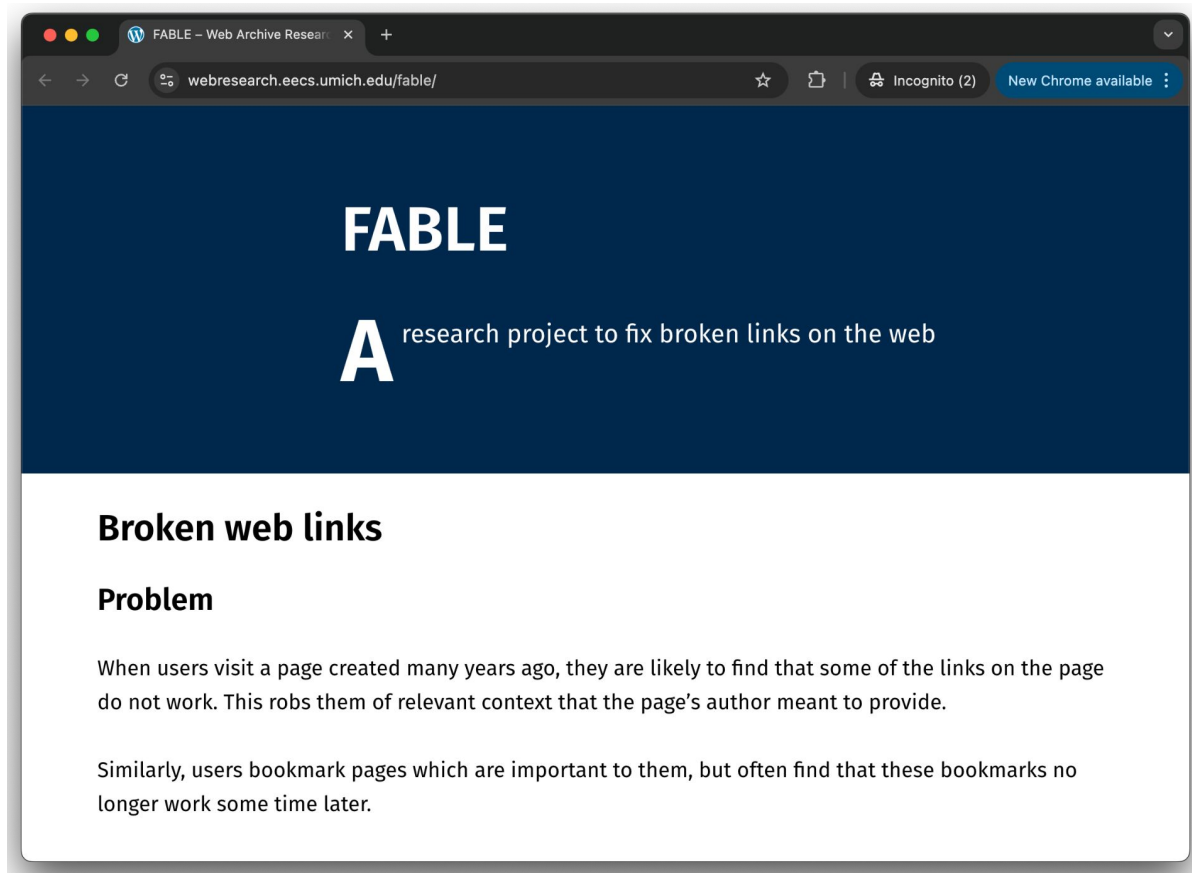
😊 Only 1 in every 10 URLs is vanished! – Wayback Machine



Vanished web: Discover new locations for the same content

- Search engine queries
- Historical redirects
- URL change patterns

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



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

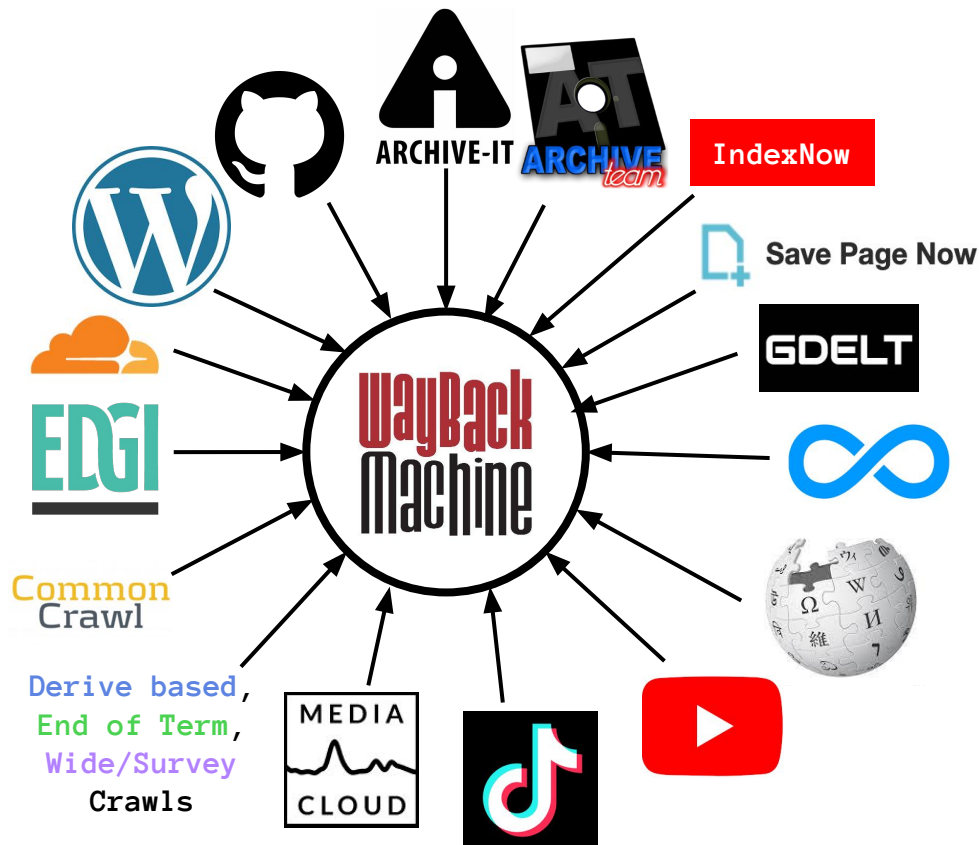
Deepweb

**Rate
limiting**

**Timely
discovery**

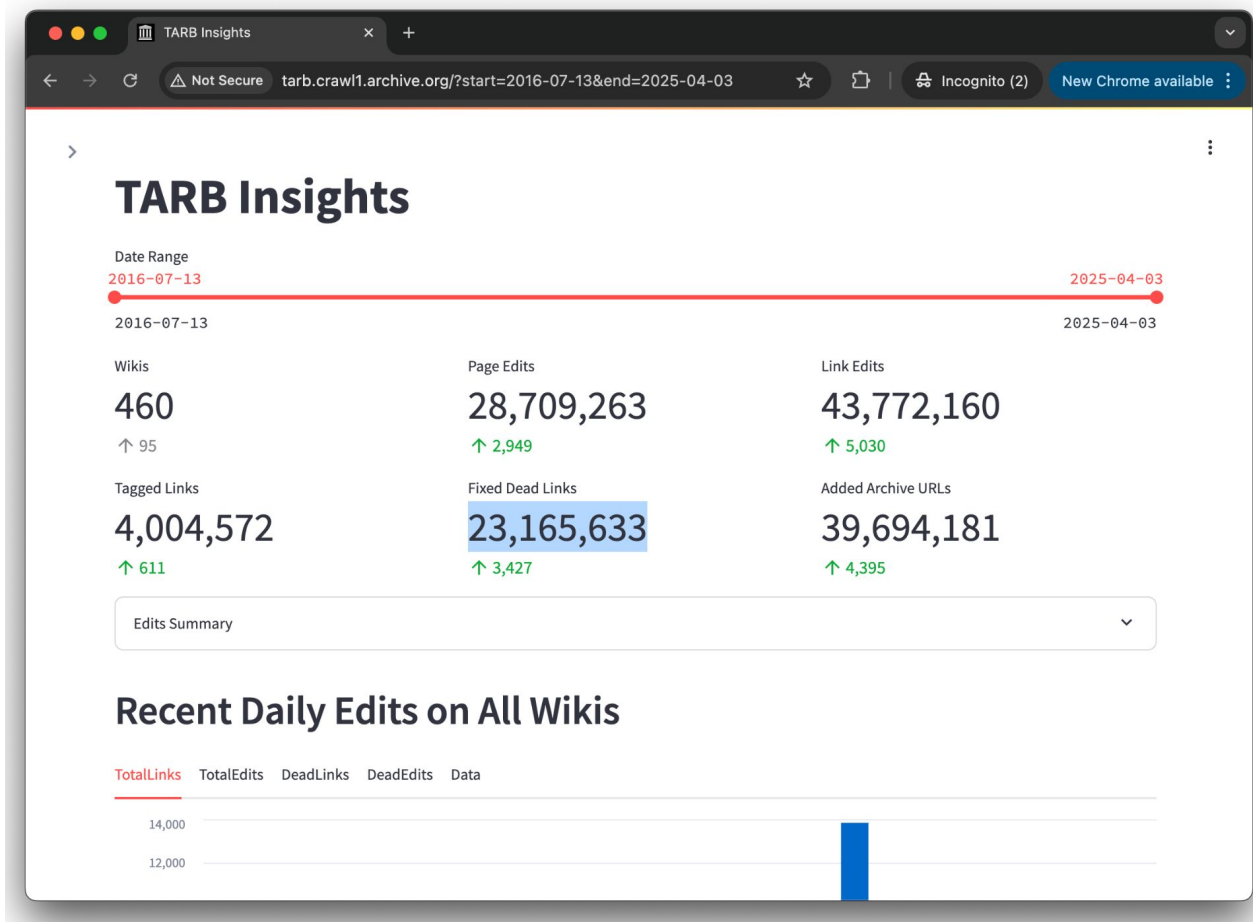
**Geo
targeting**

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!!!

TARB IABot Dataset

How many total links are there, how many are dead, how many are archived, and how many are dead but also archived?

```
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
FROM externallinks_global;
```

May I help? (e.g., How many dead links are there on the NASA page from the Portuguese wiki?)

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,
```

Run Query

	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

Collect 2013 NYTimes articles archived in the Wayback Machine.

Link extraction

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

Live status

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

Archive status

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
<i>Pew (All)</i>	2024	2013-2023	5.4M	26%	16%
<i>Pew (General)</i>	2024	2013-2023	1M	27%	13%
<i>Zittrain NYT*</i>	2021	2013-2013	88K	40%	38%
<i>ODU</i>	2024	1996-2021	27.3M	65%	65%