Lost, But Preserved

A Web Archiving Perspective on the Ephemeral Web

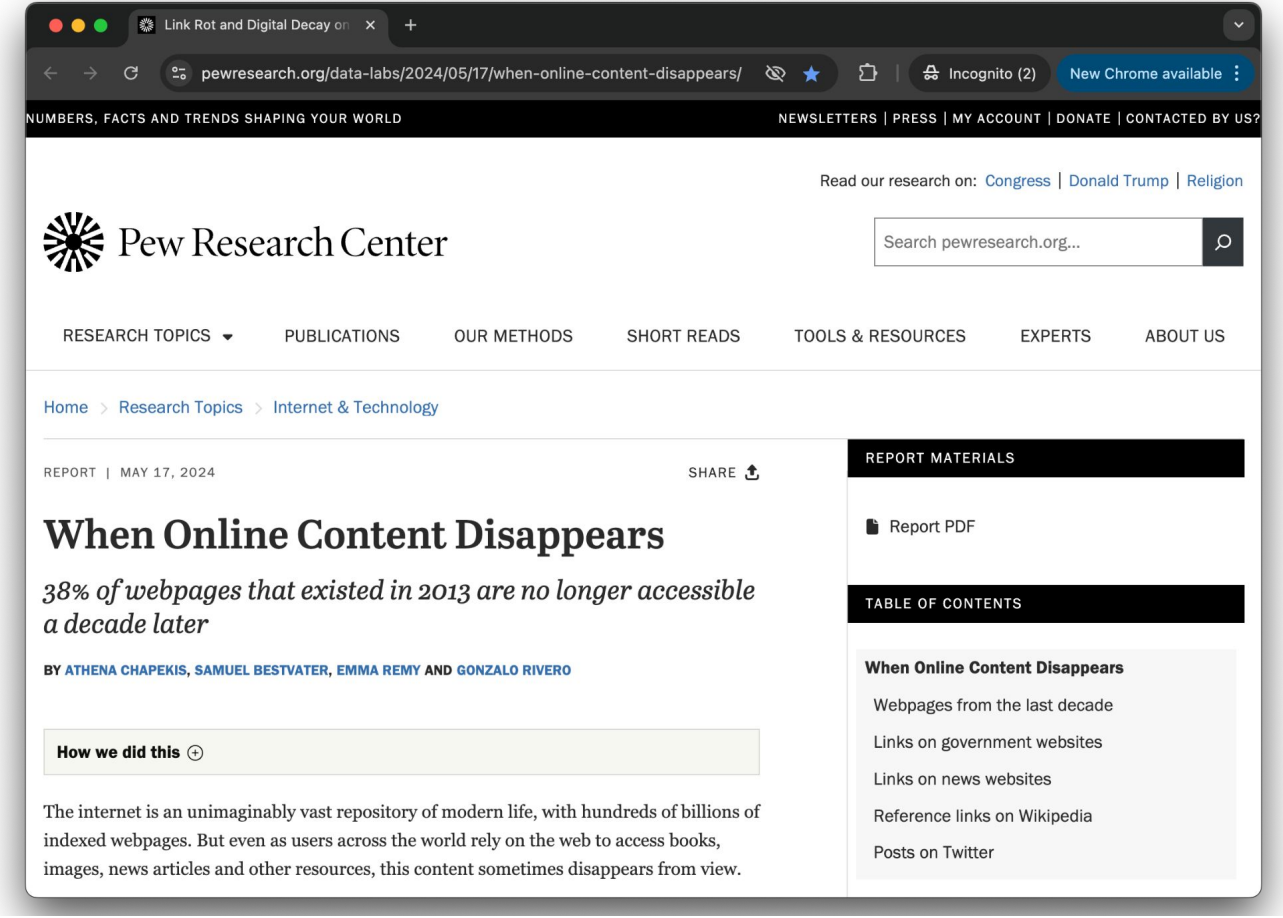
Sawood Alam & Mark Graham
Wayback Machine, Internet Archive

@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"

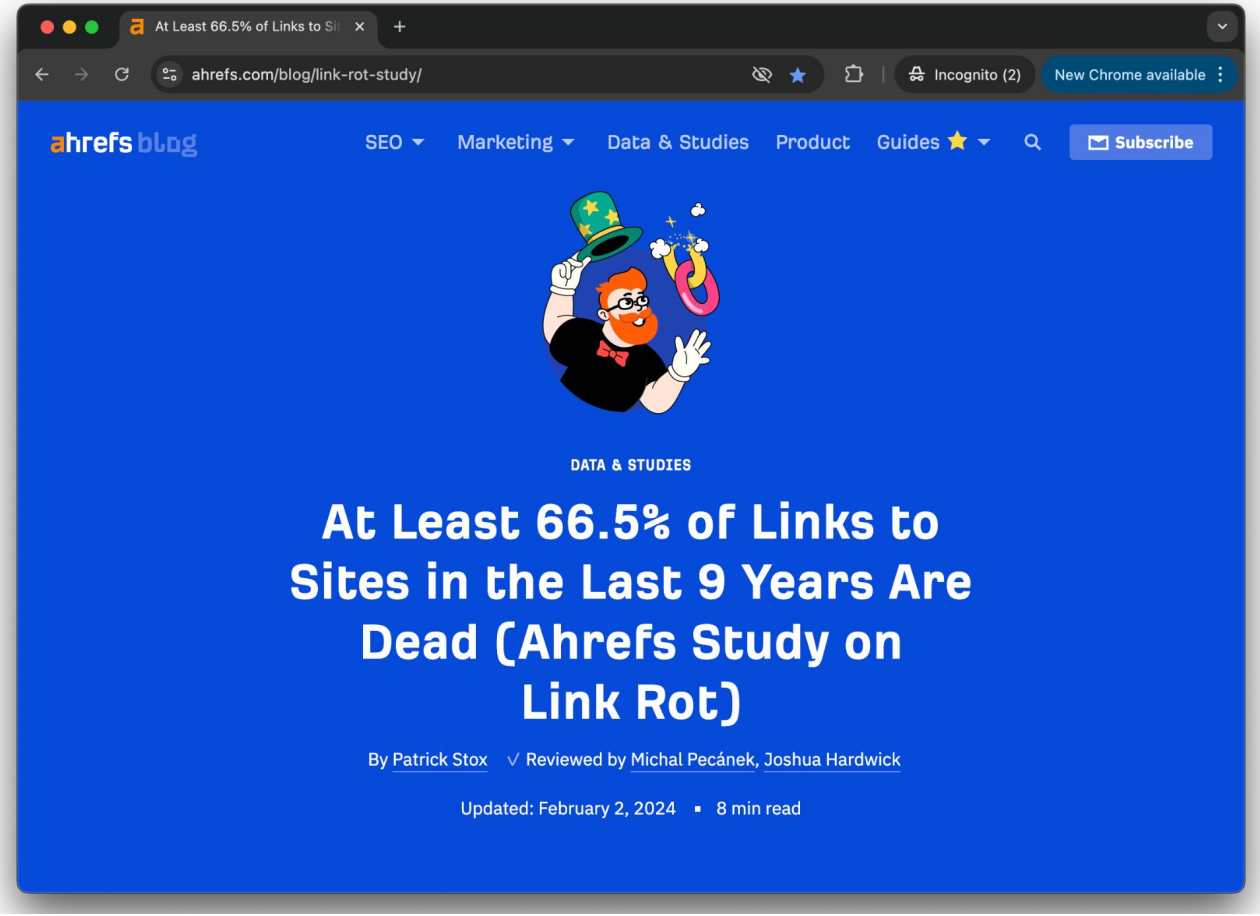
– Pew Research



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”

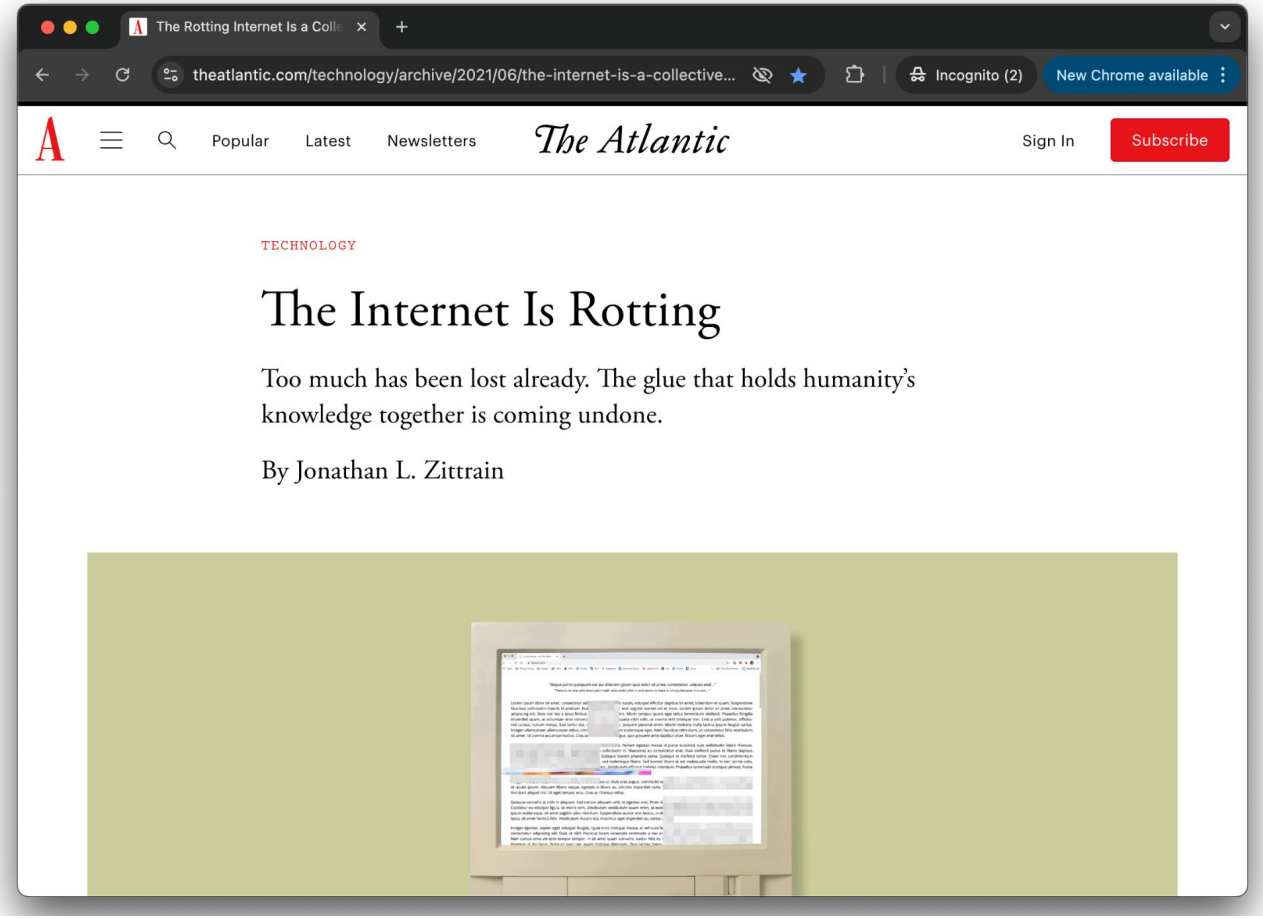
– Ahrefs



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

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

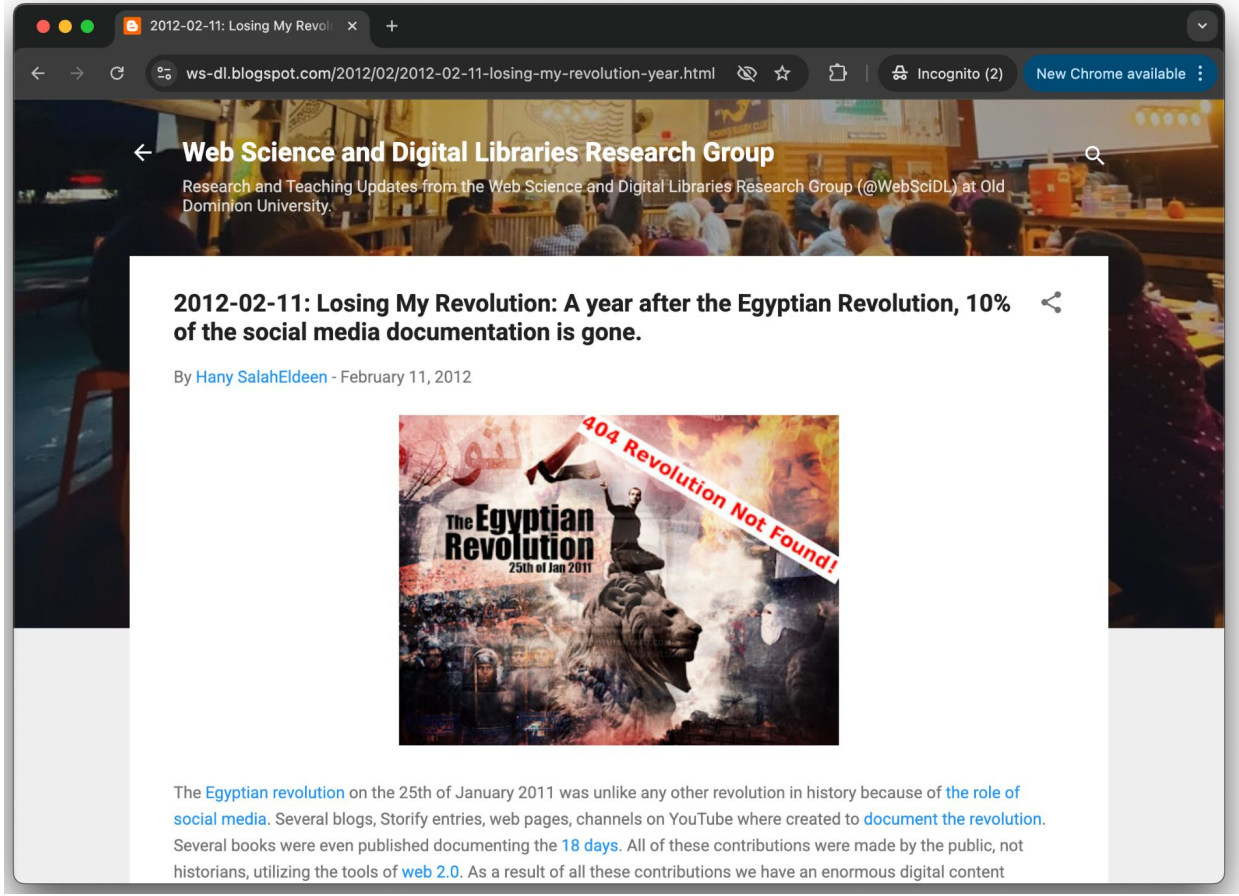
– Jonathan Zittrain



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”

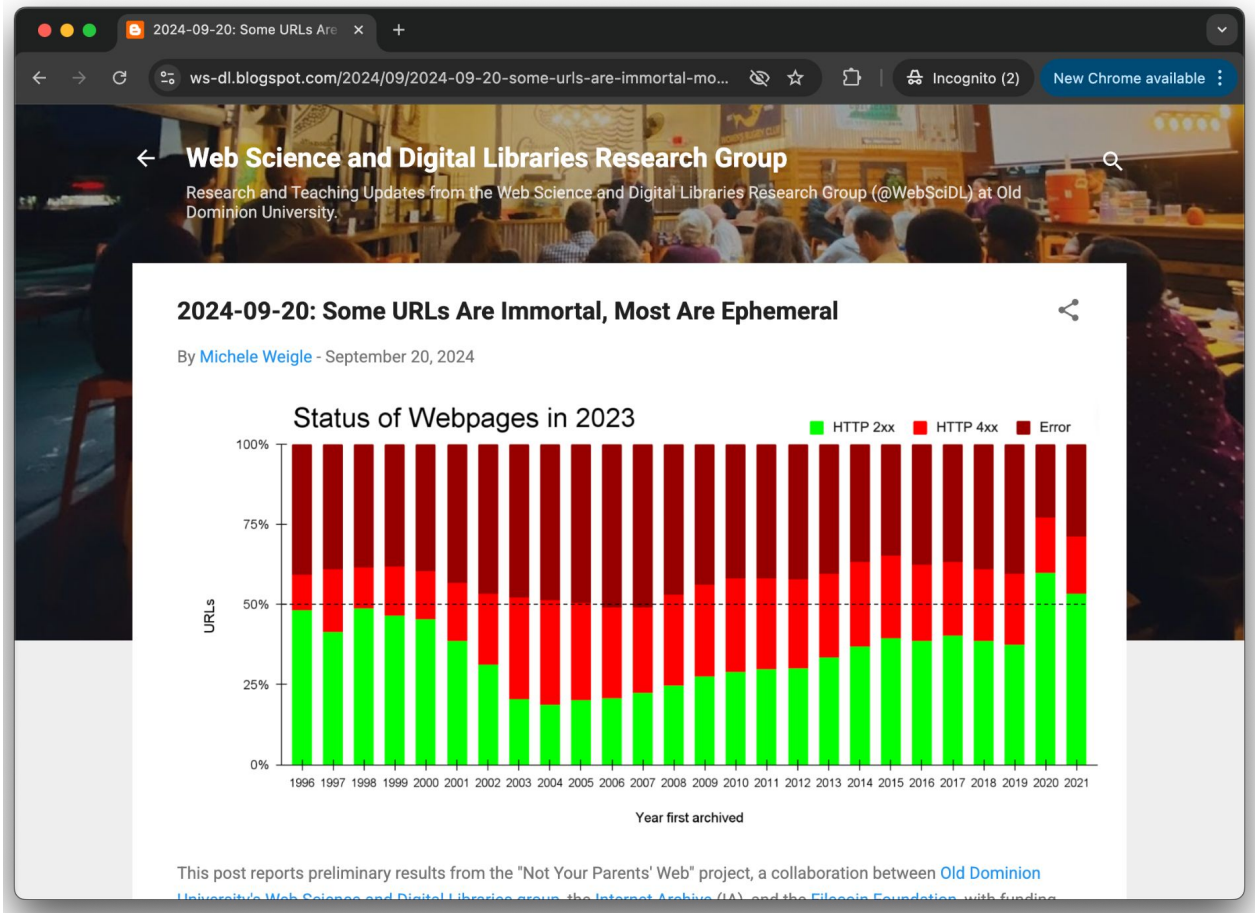
– Hany SalahEldeen



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"

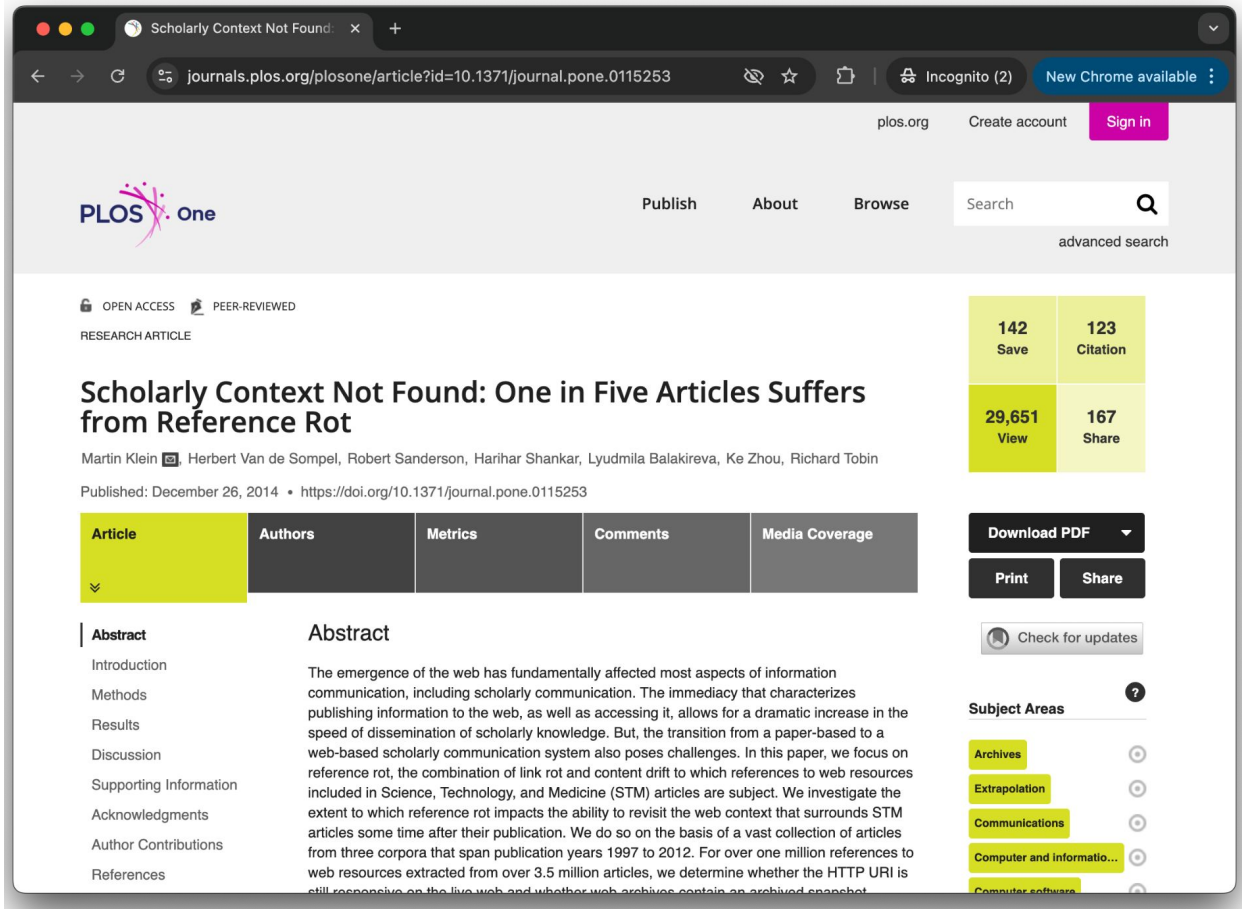
– WSDL ODU



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.



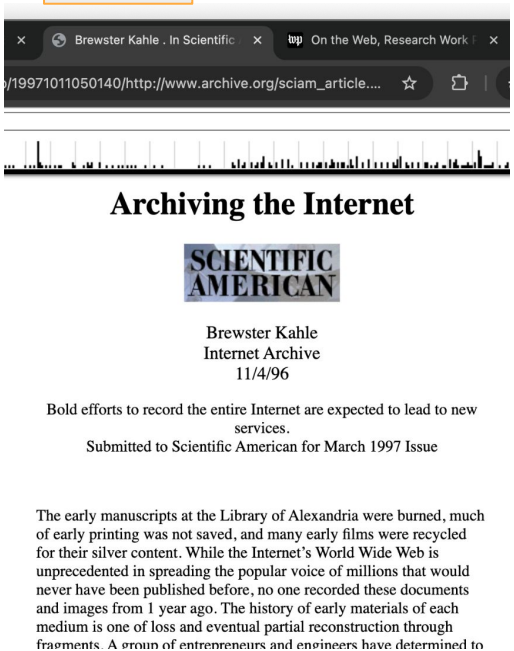
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253

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

1996

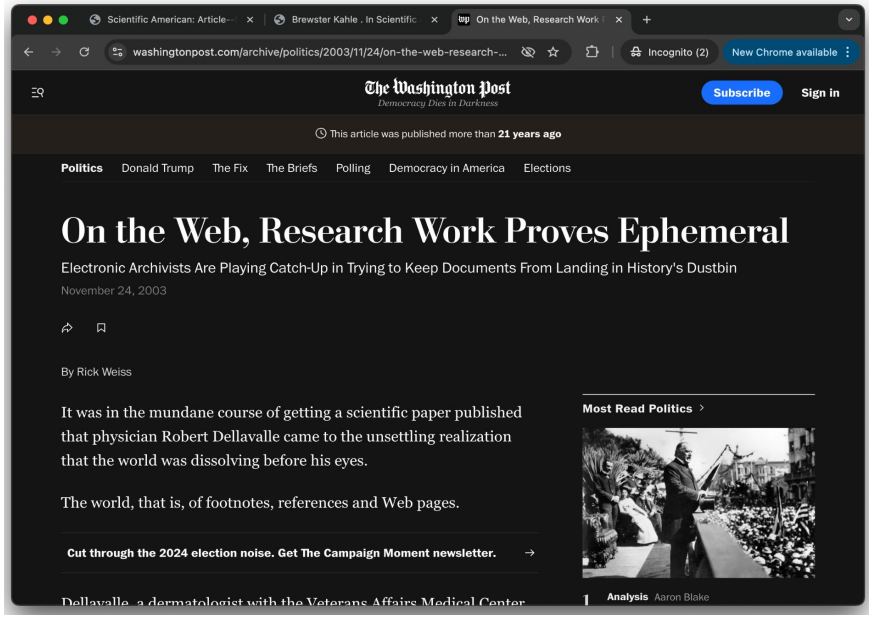
http://www.sciam.com/0397issue/0397kahle.html

1996



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

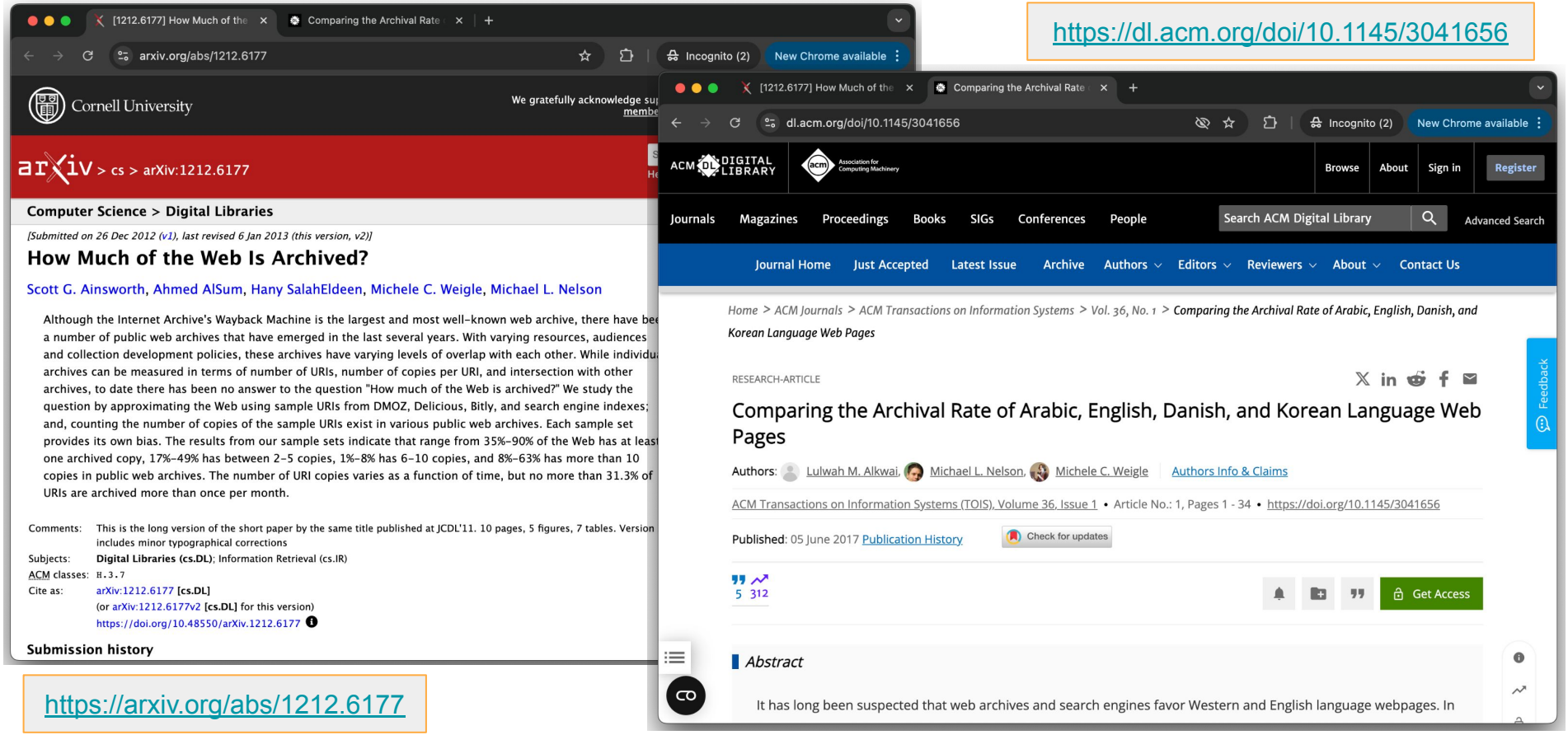
2003



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

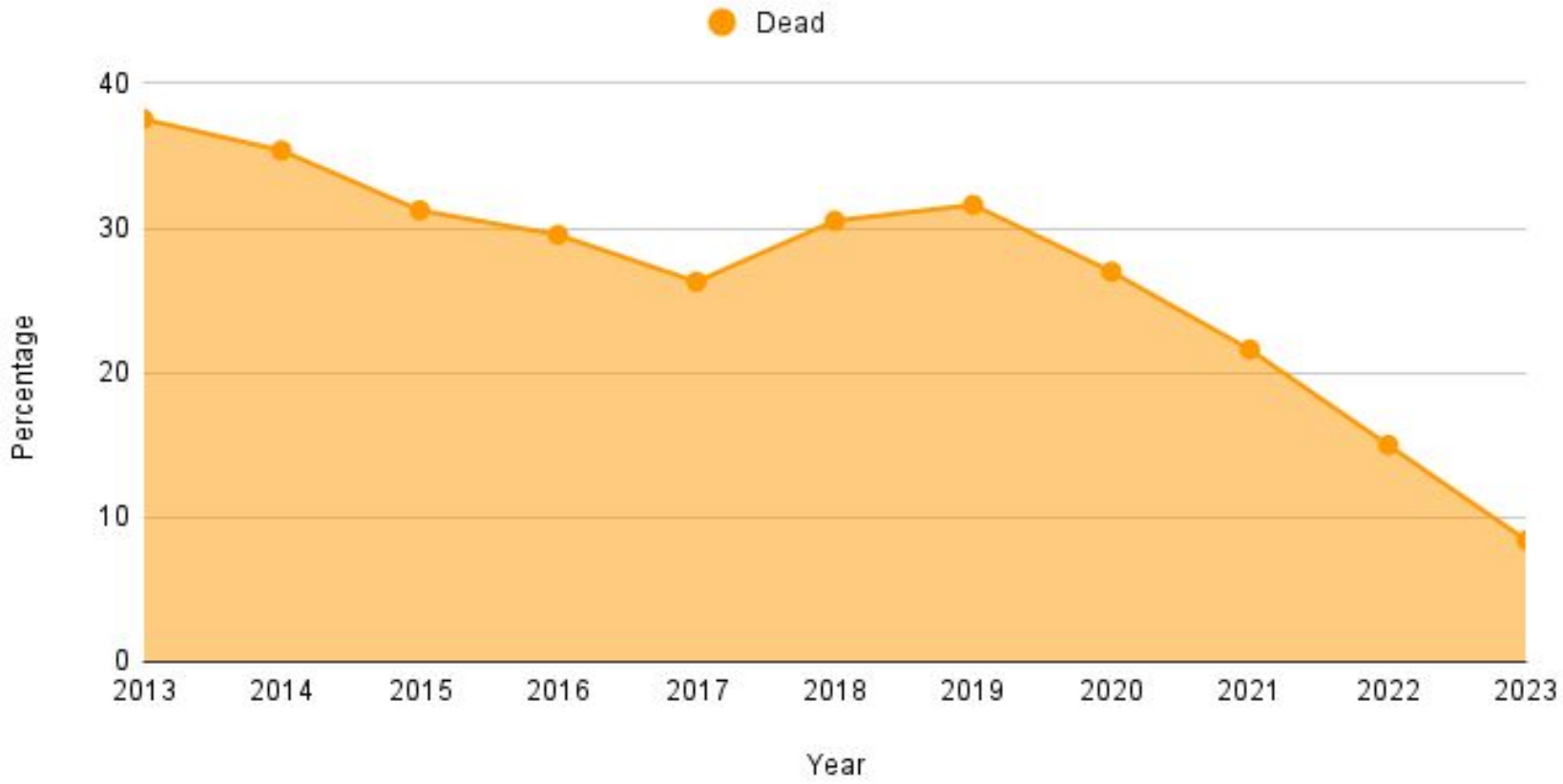


https://arxiv.org/abs/1212.6177

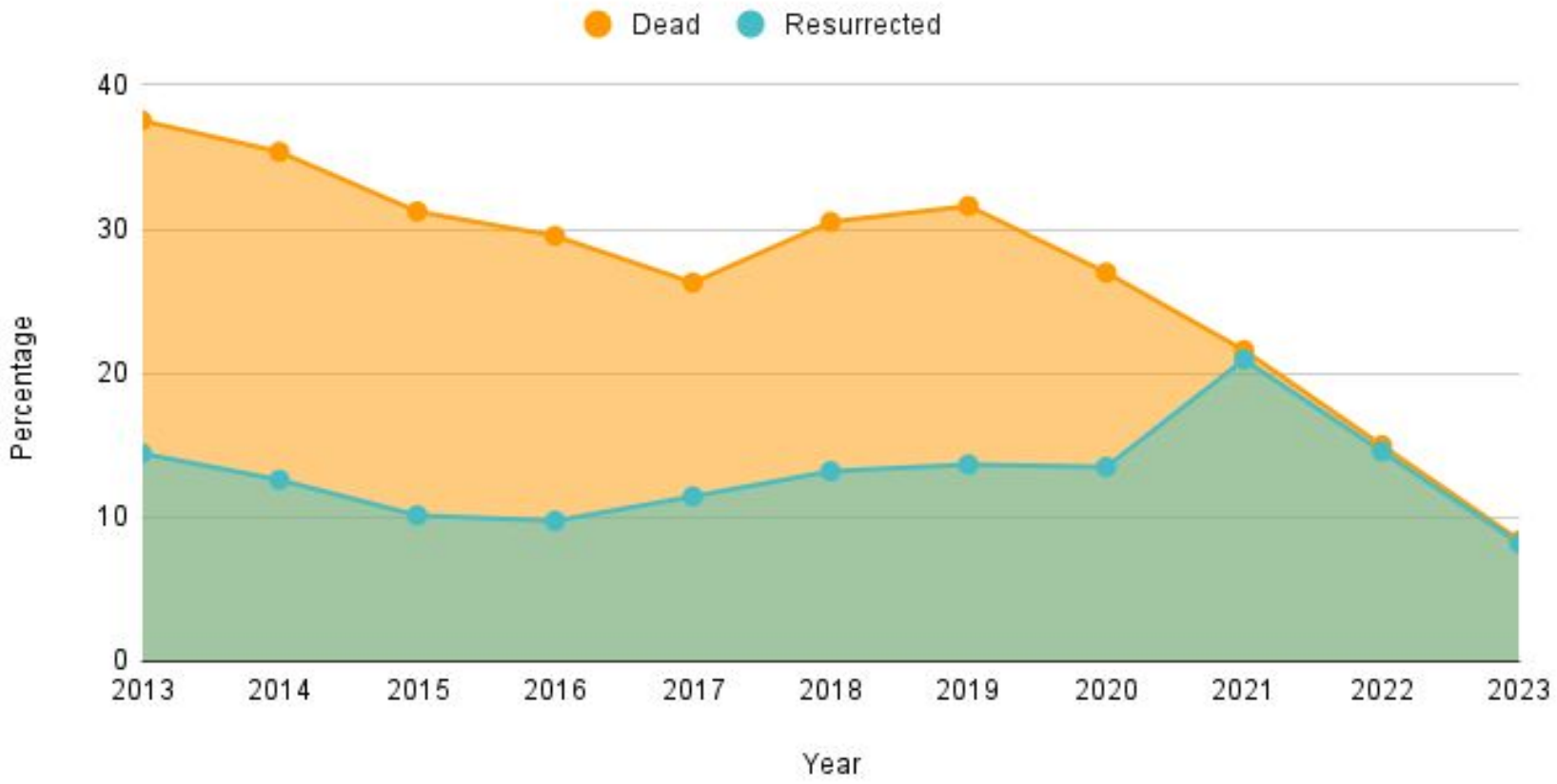
Link-rot terminologies

- ***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*

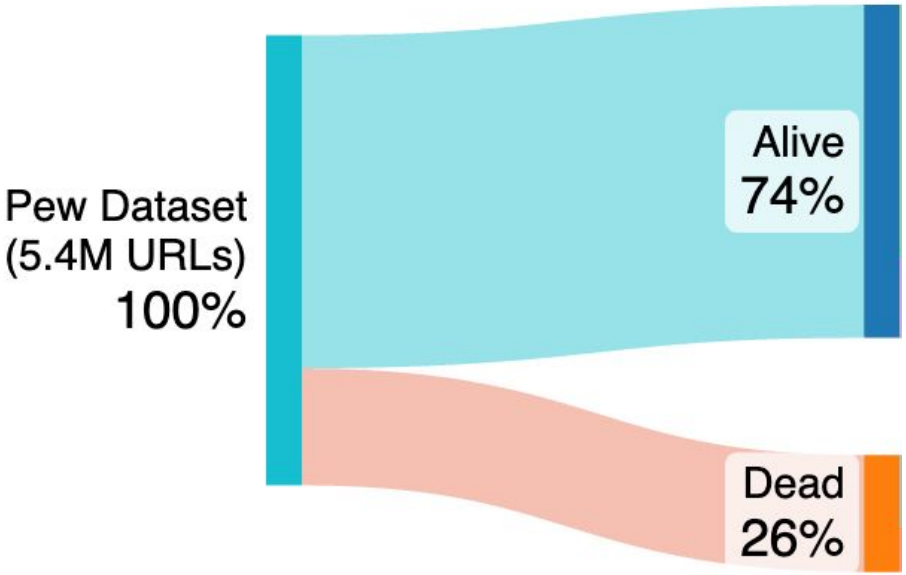
☹ 38% of the links from 2013 are dead! – Pew Research



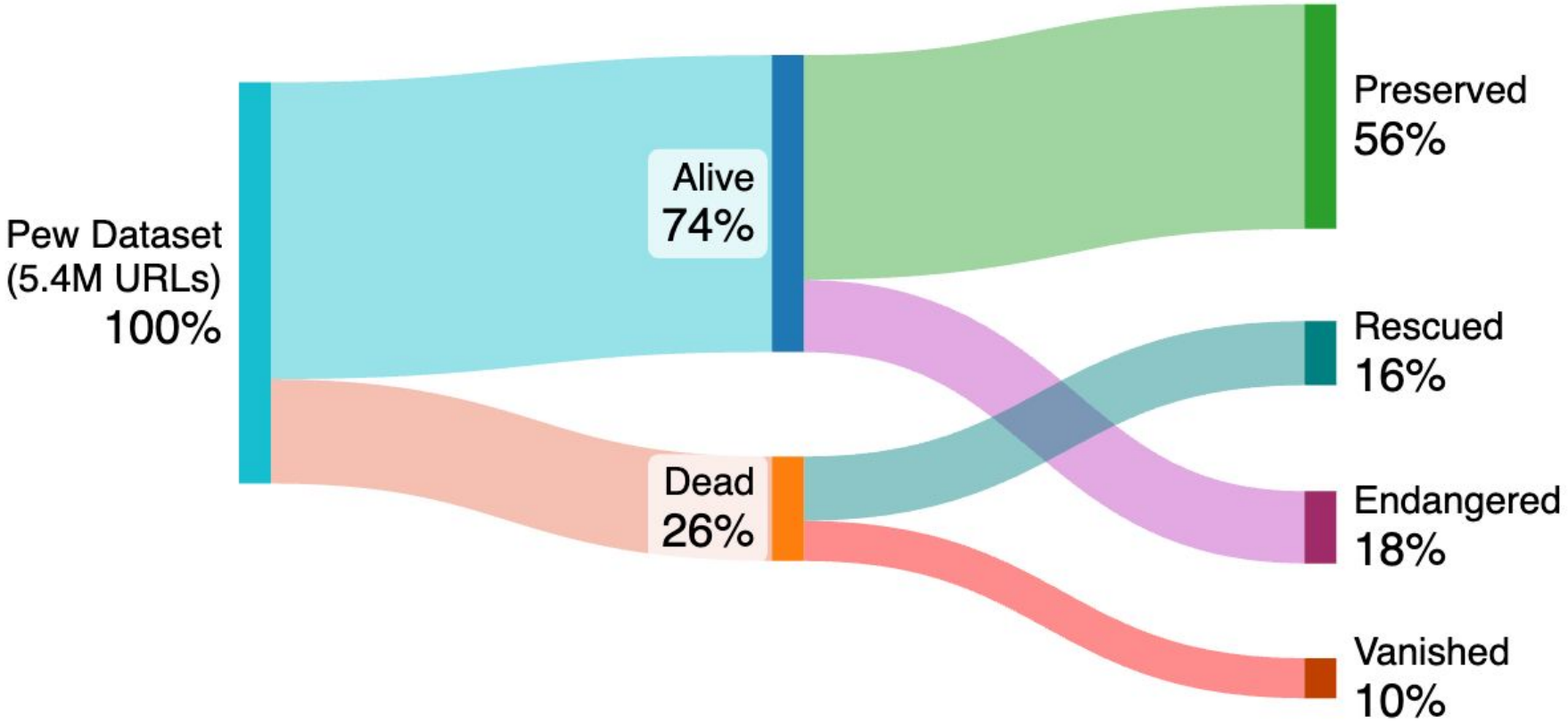
🙂 But 38% of those are resurrected! – Wayback Machine



☹ Every 1 in 4 URLs is dead! – Pew Research



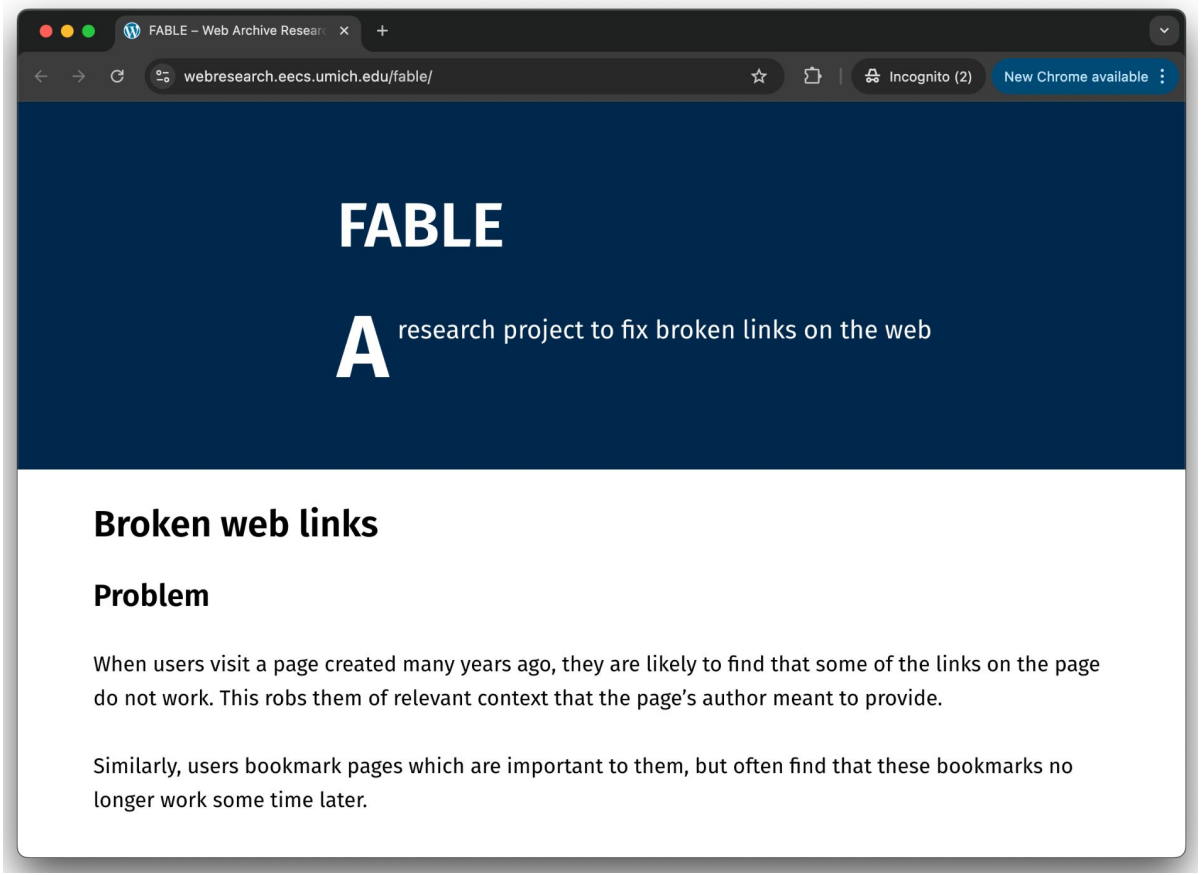
🙂 Only 1 in every 10 URLs is vanished! – Wayback Machine



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

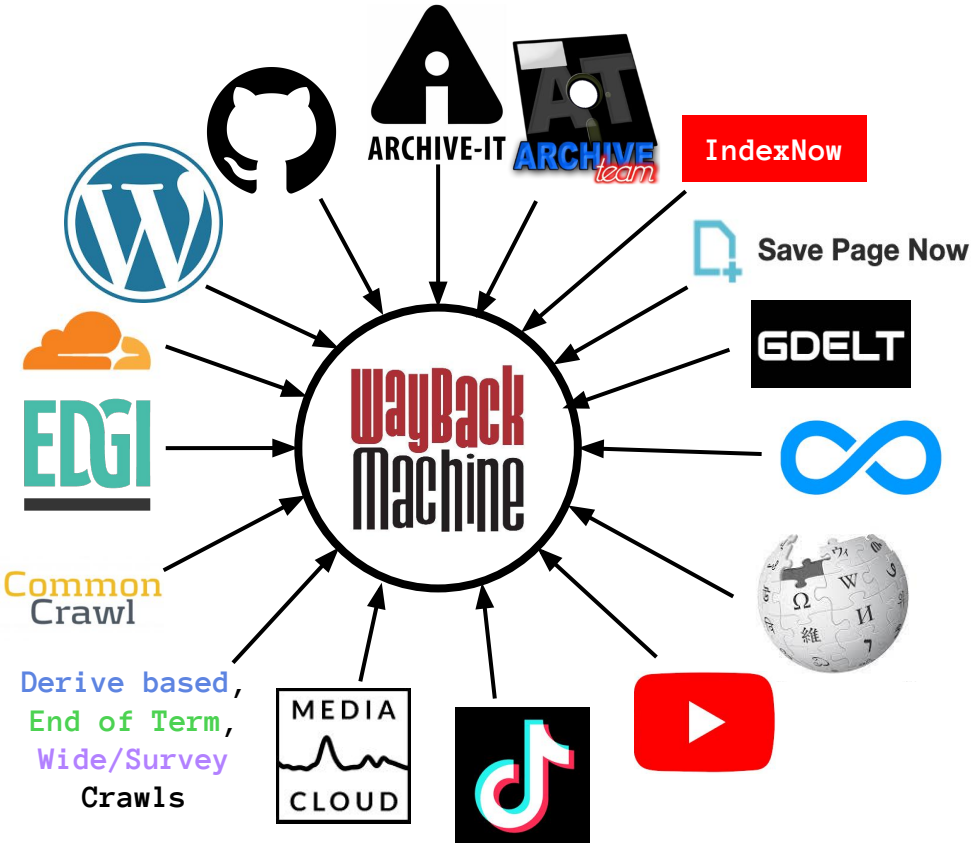
```
$ memgator -f cdxj http://si.edu/ | grep -v "^!" | cut -d'/' -f3 | sort | uniq -c | sort -nr
  13263 web.archive.org
   3590 wayback.archive-it.org
   1202 web.archive.bibalex.org
    651 webarchive.loc.gov
    321 arquivo.pt
     32 wayback.vefsafn.is
     11 web.archive.org.au
      3 archive.is
      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

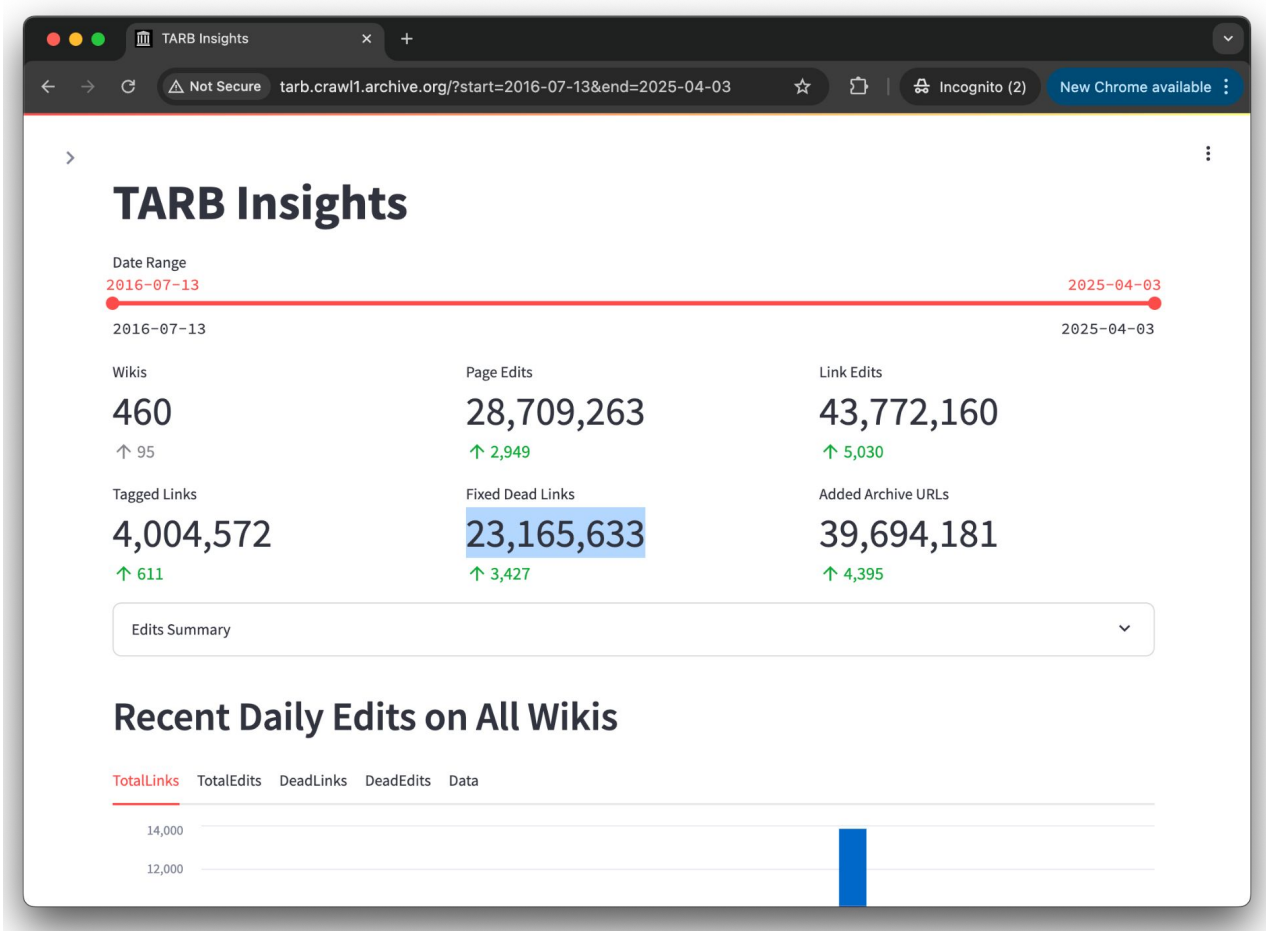
Resource constraints	**JS-heavy webpages**	**Bot blocking**
Loginwalls	**Paywalls**	**Deepweb**
Rate limiting	**Timely discovery**	**Geo targeting**

Endangered web: Many sources of link discovery & archiving



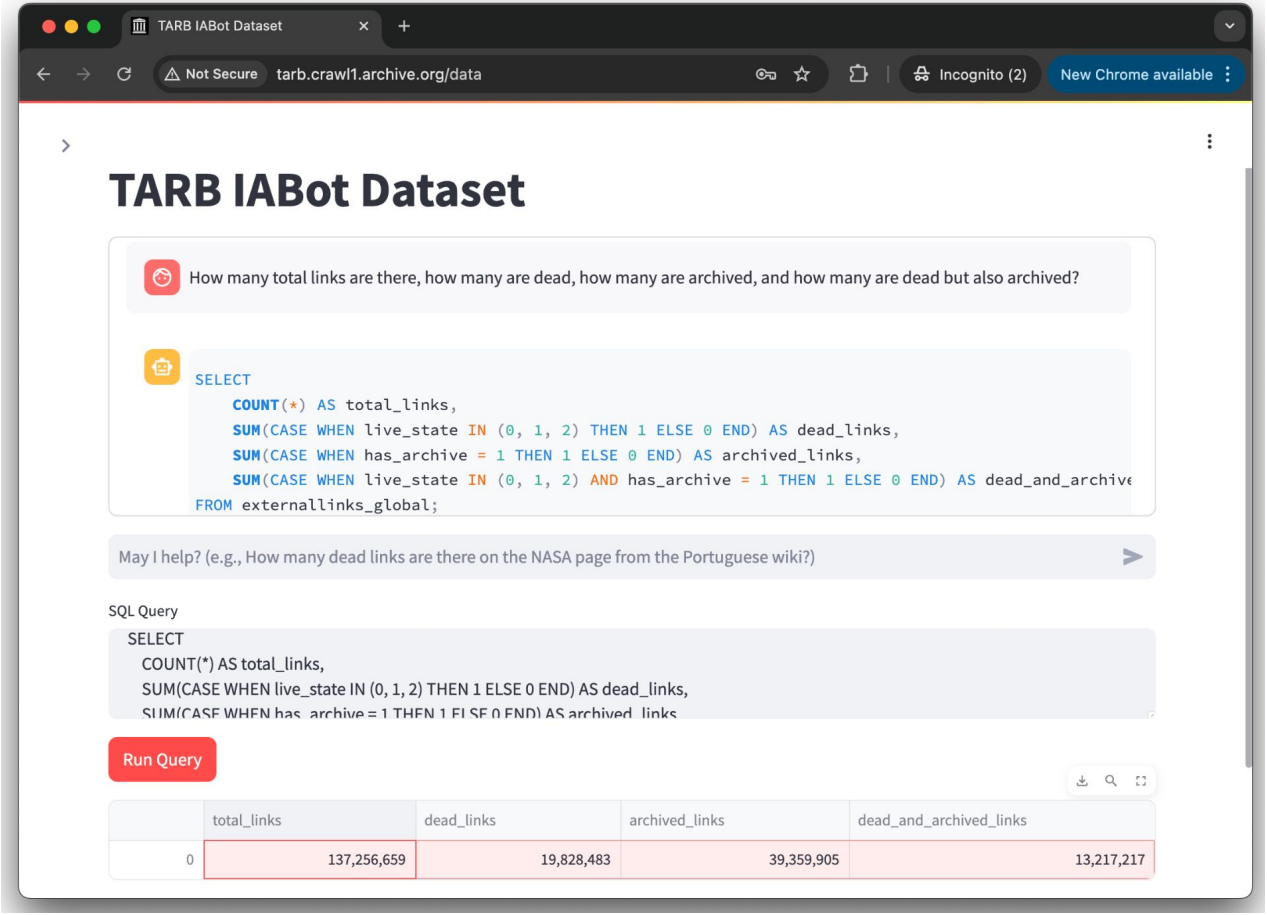
Over 1.8 billion URLs are added to the Wayback Machine every day

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



Two thirds of the dead links on Wikipedia are rescued?

Unreliable!!!



Recreating Zittrain's NYTimes dataset

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%