Table 5. Summary of links that are no longer available, broken down by link type.

Link Type	Links that Fail to Resolve	Total Links of that Type	Percent of Fail Links of that Type
URL	79	1342	5.9%
DOI	13	744	1.7%
FTP	5	21	23.8%
ZIP	7	16	43.8%
PDF	7	12	58.3%
DOCX	2	9	22.2%
GZ	3	7	42.9%
XLSX	0	5	0.0%
Badly formatted URL	1	3	33.3%
TXT	0	3	0.0%
DOC	0	2	0.0%
IMG	0	1	0.0%
ZIPR	1	1	100.0%

https://doi.org/10.1371/journal.pone.0304781.t005

files were most likely to no longer be available (58.3% loss). The PDF loss rate is complicated by the fact that many of these links point to supplementary information instead of shared datasets; this means the link information was recorded erroneously in CaltechAUTHORS (supplementary information files should have been uploaded to the repository instead of recorded as supplemental data links). As an estimated quarter of Caltech-authored articles contain supplementary information, a significant amount of which is formatted as a PDF, the continued availability of PDF data on the internet merits further study for better quantification.

How does data availability trend over time? (RQ2.2)

Examining links that were unavailable by article publication year (see Fig 2), a pattern emerges where older links were more likely to be unavailable than newer links. The number of links tested varied between publication years–with fewer supplemental data links tested for older publications–so only data from 2014–2022 is plotted and fit in Fig 2, as these years had at least 10 links to test. Modeling the relationship between availability of the datasets and age of the article using a Poisson regression, the odds of data being available was found to reduce by 2.6% (odds ratio = 0.974 [.950-.998, 95% CI], $p < 0.05$) for each year after the article is published.

Are there specific sites where data is more likely to disappear? (RQ2.3)

Tables 6 and 7 list websites and repositories, respectively, where at least two datasets have gone missing. ScienceDirect was the website with the most missing data; half of ScienceDirect's 14 links were missing. Many other journal websites appear in Table 6, suggesting that journal-hosted supplementary information is not a stable method of data sharing. As supplementary information is so prevalent (see previous discussion about PDF Links) and was only incidentally tested in this analysis, the continued availability of journal-hosted supplementary information merits further investigation. GitHub also appears at the top of Table 6, with 5 missing datasets, suggesting that it is also not an ideal place to store data in a sustainable way.

Discussion

This research has several limitations. First, this analysis is not a full measure of "data rot" as it did not check for content drift (i.e. that the harvested webpage actually represents the data

Attribution and changes

This separate note was added by Scanfile and is not part of the article.

Original: Kristin A. Briney (2024), Measuring data rot: An analysis of the continued availability of shared data from a Single University, PLOS ONE 19(6): e0304781. DOI: 10.1371/journal.pone.0304781.

Original article: https://doi.org/10.1371/journal.pone.0304781

License: Creative Commons Attribution 4.0 International (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/

Changes: selected page 9, removed interactive hyperlinks, and added this note.

The excerpt begins and ends mid-context; consult the full original article.

The author and PLOS do not endorse Scanfile or its translations.

Translation output may change wording, figures, tables and page layout.

Compare with the original and obtain expert review when accuracy matters.