Dead Science: Most Resources Linked in Biomedical Articles Disappear in Eight Years

Tong Zeng[1,2], Alain Shema[1], and Daniel E. Acuna[1](✉)

[1] School of Information Studies, Syracuse University, Syracuse, USA
deacuna@syr.edu

[2] School of Information Management, Nanjing University, Nanjing, China

Abstract. Scientific progress critically depends on disseminating analytic pipelines and datasets that make results reproducible and replicable. Increasingly, researchers make resources available for wider reuse and embed links to them in their published manuscripts. Previous research has shown that these resources become unavailable over time but the extent and causes of this problem in open access publications has not been explored well. By using 1.9 million articles from PubMed Open Access, we estimate that half of all resources become unavailable after 8 years. We find that the number of times a resource has been used, the international (int) and organization (org) domain suffixes, and the number of affiliations are positively related to resources being available. In contrast, we found that the length of the URL, Indian (in), European Union (eu), and Chinese (cn) domain suffixes, and abstract length are negatively related to resources being available. Our results contribute to our understanding of resource sharing in science and provide some guidance to solve resource decay.

1 Introduction

Reproducibility and replicability are key components of science. Increasingly, this depends on the ability of scientists to use the resources shared in scientific articles. Many studies have found that resources embedded in scientific publications suffer from decay over time [1–6] directly affecting the incremental nature of science. In particular, biomedical sciences is a discipline that reuses resources regularly (e.g., software [7], protocols [8], and datasets [9]). However, a systematic study of the decay of such resources in biomedical publications is lacking.

The mechanisms governing sharing of data and resources are important for science. As early as 2003, the National Institutes of Health published a policy requiring applications for grants greater than $500,000 to include data sharing plans [10]. The National Science Foundation also has policies encouraging data sharing [11]. There are other institutions that recognize the importance of this practice (e.g., [12,13]) and its actual impact on the acceleration of science

N. G. Taylor et al. (Eds.): iConference 2019, LNCS 11420, pp. 170–176, 2019.
https://doi.org/10.1007/978-3-030-15742-5_16

(e.g., [9]). Sharing of resources is important and how they decay is still poorly understood.

One way of understanding how long and why resources are available is to analyze how resources decay over time. Several studies have tried to understand this phenomenon in several disciplines. In [14], the authors extracted 4,387 unique URLs published from 1995 to 2004 in D-Lib Magazine, revealing that 30% of the URLs failed to resolve, leading to a half-life of approximately 10 years. A similar quantity of 9.3 years was found by [15] using median survival lifespan analysis. There are similar studies in Law [16], Ecology [3] and Library and Information Science [4,14]. [2] analyzed 2,822 medical articles in a small sample of journals, finding half-lives between and 2.2 and 5.3 years. All this previous work has focused mostly on closed access publications and, to the best of our knowledge, the biggest dataset has around 1 million URLs [5,6]. In our work, we examine resource decay at a significantly larger volume in open access biomedical articles.

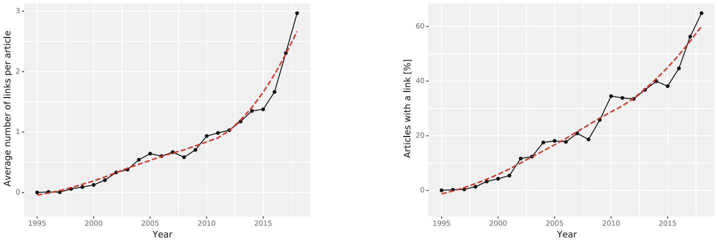
2 Materials and Methods

We obtained a copy of Pubmed Open Access Subset in June 2018 which consists of 1,904,971 articles. Not all URLs in these files are interesting or represent a resource being shared. We apply the following filters to discard URLs. First, we remove links to local file systems, URLs without any paths, and we canonicalize the URLs. The URL availability checker followed standard detection methods [17]. This checker, however, does not consider resources that are available but moved from the original URL. The final dataset contains 2,642,694 URLs of which 1,883,622 are unique.

3 Results

Exponential Growth in Link Sharing. We wanted to examine how resource sharing in the form of URLs has evolved over time. For each year of publication, we computed the number of links per article. This trend is exponential (Fig. 1a). We also analyze the percentage of articles with at least one URL. We found that this trend is exponential and that articles published today are more likely than not to have a link to a resource (Fig. 1b). These findings suggest that a URL has become a primary mechanism for sharing resources and that keeping track of decay of these resources will therefore become more important.

Most Resources Shared in Publications Disappear After Eight Years. The point at which half of the resources become obsolete is important. Here we simply examined the average availability of a resources as a function of age (Fig. 2) and found that this point happens after eight years. Surprisingly, we notice that new resources (age = 0) have a 20% chance of being unavailable, similar to previous findings [18]. Resources tend to follow a steady decline from ages 1 to 10 years. Then, it seems that resources 10 years and older stabilize around 42% availability. The data in our research showed that half of links



(a) Average links per paper as a function publication year. A locally weighted regression (loess, red line) is shown as well..

(b) Percent of articles with a link as a function of publication year. A locally weighted regression ("loess", red line) is shown as well.

Fig. 1. Some trends about the resource sharing. There is a clear exponential growth in these trends. (Color figure online)

become unavailable after 8 years, but this trend does not continue on the second 8-year period, as the available percentage fluctuates around 42%. Thus, the decay trend does not follow the typical half-life analysis found in Physics research [19].

Factors Related to Link Availability. The availability of resources might be associated with several characteristics. We computed several features to get at these characteristics (Table 1). Some of these features are related to the article (e.g., h-index of the journal, number of authors and affiliations, title), and the link itself (e.g., number of occurrences, link length). Also, we analyzed the suffixes of the domains to get at country or organizational effects. These features allowed us to tell apart several factors influencing link presence.

We used a logistic regression model with elastic net regularization [20] on standardized features to understand the relative importance of all the previously described factors. The cross validated performance of our model has an $F_1 = 0.67$. The most positive weights were the number of times a link has been used and the domain suffix "int". Interestingly, the "org" domain suffix, the size of the resource being shared, and the number of affiliations in the paper had all large positive associations. These positive associations intuitively suggest that links that have been shared in many articles, articles in government or non-profit organizations, and articles with many authors are all factors that contribute to a link being available.

There are other features that are negatively associated with availability. Expectedly, the most important negative feature is the link's age. In absolute terms, it is also the biggest contributor to the prediction. The length of the path (related to the length of the URL) is also an important predictor. Domain suffixes related to India (in), the European Union (eu), China (cn), and Korea (kr) are all also negatively related to availability (Table 1). These features could help to identify which links are in danger of becoming lost.

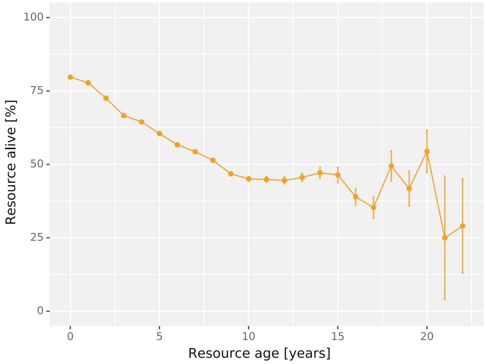


Fig. 2. Probability of resources being available as a function of age in years.

Table 1. List of top features and their standardized weights for predicting availability in a regularized logistic regression model

Features	Feature type	Weight
The frequency of a URL across articles	Link	0.26
int (domain suffix)	Link	0.20
org (domain suffix)	Link	0.12
gov (domain suffix)	Link	0.07
Size in bytes of the resource referred by the URL	Link	0.07
Number of affiliations of the article which cites the URL	Article	0.07
The h-index of a journal	Article	0.02
Years back from 2018	Article	−0.37
The length of the path in the URL	Link	−0.18
in (domain suffix)	Link	−0.06
eu (domain suffix)	Link	−0.05
cn (domain suffix)	Link	−0.05
The number of query string parameters in the URL	Link	−0.05
Length of abstract of the article	Article	−0.05
Number of references of the article	Article	−0.03
The length of the query string in the URL	Link	−0.02
Number of authors of the article which cites the URL	Article	−0.01
Length of article’s title	Article	−0.01

Top Cited Links Are Mostly Tools. We performed a qualitative analysis of the most shared links and their availability. Interestingly, most of these highly cited links were tools related to gene expression and sequence search. The most cited of them is the Gene Expression Omnibus (GEO) from NIH. Several URL aliases related to the image processing tool ImageJ were at the top. Japan’s

Kyoto Encyclopedia of Genes and Genomes and UK's figtree were two top non-US tools. Conversely, a very popular tool cited over 600 times was unavailable at https://tcga-data.nci.nih.gov/publications/tcga. This suggests that tools tend to be more available because they are used by many people, are maintained by a team, and are required for reproducibility. These observations could be applied to other non-tool resources to make them last longer.

4 Discussion and Conclusion

The practice of embedding links in scientific papers has been growing exponentially, and our findings are in line with previous research [3,15]. While we found a half-life of 8 years, there is great variability in this number—2.2 years [2], 5 years [4], 5.3 years [2], 9.3 years [15], 10 years [14]. However, all this previous work has analyzed a smaller volume of links and shorter time spans compared to our analysis, which may explain this variability.

There could be other more detailed analysis of how links become unavailable. Unavailability is mainly due to two types of problems: (1) the URL becomes inaccessible or (2) the content of the resource changes since the publication—a phenomenon known as content drift [21]. The first type of problem is usually related to change of domain, movement of resource, or cessation of operation. The second type of problem can be detected by manual checking [16] or using third-party web archiving services [6]. Due to the dynamic nature of the web, tackling content drift at scale is a challenging.

There are few research projects aimed at modeling link decay in science. [21] builds several SVM models to predict the availability of a link, achieving a performance of 0.72 in AUC. SVM classification analysis are however hard to interpret. [15] uses survival regression modeling to predict the median survival lifetime of a link and provide interpretable results. However, survival analysis does not predict probability of availability. Our long term-goal is to predict availability and understand the factors affecting it, and our logistic regression achieves these two goals simultaneously. If only performance is a concern, we will in the future explore more advanced techniques, such as deep learning, that sacrifice interpretability for better predictions.

While our findings are only correlations, they offer some intuitive suggestions. We would propose that authors use shorter, easier to remember URLs, hosted in non-profit domains. For links that are available, we should consider archiving those that are old, have complex URLs, are published in low h-index journals, and are hosted in country-based domains. In none of this is possible, we can explicitly archive links with services such as Perma [16], the Internet Archive [22], and WebCite [23,24].

In spite of the shortcomings of the present research, it offers previously unknown factors affecting link decay in open access biomedical journals at a much larger scale than before. Future research will investigate how to develop more precise predictions by expanding the data sources beyond open access biomedical articles, by improving the URL checking process, and by increasing the complexity of prediction model.

Acknowledgements. Tong Zeng was funded by the China Scholarship Council #201706190067. Daniel E. Acuna was funded by the National Science Foundation awards #1646763 and #1800956.

References

1. Koehler, W., et al.: A longitudinal study of web pages continued: a consideration of document persistence. Inf. Res. **9**(2) (2004)
2. Habibzadeh, P.: Decay of references to web sites in articles published in general medical journals: mainstream vs small journals. Appl. Clin. Inf. **04**(4), 455–464 (2013)
3. Duda, J.J., Camp, R.J.: Ecology in the information age: patterns of use and attrition rates of internet-based citations in ESA journals, 1997–2005. Front. Ecol. Environ. **6**(3), 145–151 (2008)
4. Goh, D.H.-L., Ng, P.K.: Link decay in leading information science journals. J. Am. Soc. Inf. Sci. Technol. **58**(1), 15–24 (2007)
5. Klein, M., et al.: Scholarly context not found: one in five articles suffers from reference rot. PloS ONE **9**(12), e115253 (2014)
6. Jones, S.M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., Grover, C.: Scholarly context adrift: three out of four URI references lead to changed content. PLoS ONE **11**(12), e0167475 (2016)
7. Mangul, S., et al.: A comprehensive analysis of the usability and archival stability of omics computational tools and resources. bioRxiv, p. 452532 (2018)
8. Collaboration, O.S., et al.: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015)
9. Bonàs-Guarch, S., et al.: Re-analysis of public genetic data reveals a rare x-chromosomal variant associated with type 2 diabetes. Nature Commun. **9** (2018)
10. National Institutes of Health: Final NIH statement on sharing research data (2003). https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html. Accessed 5 Dec 2018
11. National Science Foundation: NSF data sharing policy (2017). https://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp#VID4. Accessed 5 Dec 2018
12. Van Horn, J.D., Gazzaniga, M.S.: Why share data? Lessons learned from the fMRIDC. NeuroImage **82**, 677–682 (2013)
13. Milham, M.P., et al.: Assessment of the impact of shared brain imaging data on the scientific literature. Nature commun. **9** (2018)
14. McCown, F., Chan, S., Nelson, M.L., Bollen, J.: The availability and persistence of web references in d-lib magazine. arXiv preprint cs/0511077 (2005)
15. Hennessey, J., Ge, S.X.: A cross disciplinary study of link decay and the effectiveness of mitigation techniques. BMC Bioinf. **14**, S5 (2013)
16. Zittrain, J., Albert, K., Lessig, L.: Perma: scoping and addressing the problem of link and reference rot in legal citations. Legal Inf. Manag. **14**(2), 88–99 (2014)
17. Gourley, D., Totty, B., Sayer, M., Aggarwal, A., Reddy, S.: HTTP: The Definitive Guide. O'Reilly Media Inc. (2002)
18. Aronsky, D., Madani, S., Carnevale, R.J., Duda, S., Feyder, M.T.: The prevalence and inaccessibility of internet references in the biomedical literature at the time of publication. J. Am. Med. Inf. Assoc. **14**(2), 232–234 (2007)
19. Burton, R.E., Kebler, R.: The 'half-life' of some scientific and technical literatures. Am. Documentation **11**(1), 18–22 (1960)

20. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
21. Zhou, K., Grover, C., Klein, M., Tobin, R.: No more 404s: predicting referenced link rot in scholarly articles for pro-active archiving. In: Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL 2015, pp. 233–236. ACM Press (2015)
22. Internet Archive: Wayback machine. https://archive.org/web/. Accessed 5 Dec 2018
23. Eysenbach, G., Trudel, M.: Going, going, still there: using the webcite service to permanently archive cited web pages. J. Med. Internet Res. **7**(5), e60 (2005)
24. Eysenbach, G.: Preserving the scholarly record with webcite(r): an archiving system for long-term digital preservation of cited webpages. In: Proceedings ELPUB 2008 Conference on Electronic Publishing, pp. 378–389, Toronto, Canada (2008). www.webcitation.org