Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04716-z OPEN

Predicting the longevity of resources shared in scientific publications

Daniel E. Acuna [1]✉, Jian Jian[2], Tong Zeng[3], Lizhen Liang [4] & Han Zhuang [5]

Research has shown that most resources shared in articles (e.g., URLs to code or data) are not kept up to date and mostly disappear from the web after some years (Zeng et al., 2019). Little is known about the factors that differentiate and predict the longevity of these resources. This article explores a range of explanatory features related to the publication venue, authors, references, and where the resource is shared. We analyze an extensive repository of publications and, through web archival services, reconstruct how they looked at different time points. We discover that the most important factors are related to *where* and *how* the resource is shared, while surprisingly little consideration is given to the author's reputation or prestige of the journal. By examining the places where long-lasting resources are shared, we suggest that it is critical to educate researchers on modern sharing technologies. Finally, we discuss implications for reproducibility and acknowledge scientific datasets as first-class citizens of science.

[1] University of Colorado at Boulder, Boulder, CO, USA. [2] NetApp, San Jose, CA, USA. [3] Virginia Tech, Blacksburg, VA, USA. [4] Syracuse University, Syracuse, NY, USA. [5] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China. ✉email: daniel.acuna@colorado.edu

Introduction

Scientific research produces not only publications but also datasets and computer code. These artifacts have become increasingly crucial for reproducibility and replication. It is concerning to see from recent research that resources shared in scientific articles tend to disappear over time. For example, a recent study estimated that more than half of the URLs pointing to web resources inside scientific articles disappear after eight years (Zeng et al., 2019). In this study, we investigate the factors that are predictive of the longevity of these resources. It is difficult to rectify the situation without understanding these factors. Hence, we propose analyzing factors such as authors, the journals, and the technology behind sharing information to understand this association. We attempt to use our research results to illuminate this underdeveloped area of scientific resources beyond publications.

Sharing resources has always been an essential part of the scientific process. For example, it can help increase the reproducibility of results of other studies (Vasilevsky et al., 2013), allowing fellow scientists to build upon previous work more efficiently. In addition, replicability is a more vital form of scientific correctness check, requiring an entirely new group to redo and confirm experimental results (National Academies of Sciences & Medicine, 2019). Thus, in all these settings, sharing resources is essential.

Historically, it has been expensive to share resources in scientific work. Also, once resources are shared, they might need to be kept available for an extended period. Several studies have revealed that resources are surprisingly often lost prematurely. For example, Zhou et al. use features extracted from both the article and URLs to predict the likelihood of a link being available as binary classification task (Zhou et al., 2015). A recent study analyzed approximately 1.88 million of URLs linking to external resources with 1.9 million articles in biomedical science. It was found that more than 50% of the resources became unavailable after eight years of its first release. Additionally, the study revealed that the frequency of URL utilization frequency, the international (int) and organization (org) domain suffixes, and the number of affiliations in the article are positively related to resources being available or not (Zeng et al., 2019). The work of Zeng et al., 2019 focuses on the measurement of the web link half-life only and the factors related to a link being available or not. While there are new methods to capture which datasets and other artifacts are used and cited in the literature (Zeng et al., 2024, Zhuang et al., 2023), there is little research on modeling how long the linked web resources will last and examining the factors predictive of it. Resource sharing in science has become more manageable, but little is understood about why their upkeep fails across time.

In this study, we have the following contributions:

1. We enriched the data from (Zeng et al., 2019), and reconstructed how the resources and publications would have looked like at the time of publication. In this manner, we know precisely *when* a resource disappeared from the web. Also, we compute features that we hypothesize are predictive of when a resource disappears, including the type of technology used for sharing, the prestige of the authors, and publication venues.
2. We built two statistical models for the prediction of resources longevity: one highly interpretable based on a censored model and one highly predictive but less interpretable based on Random Forests.
3. We performed analysis that reveals that the technology where the resources are stored is essential for predicting resource longevity. Surprisingly, we find that the prestige of the author and institution are almost irrelevant.
4. We discussed the implications of these results related to how we can plan to keep resources available in the future.

Literature review

A wide range of different types of contents are shared in research articles, to name a few: figures, tables, formulas, andtexts. These resources are not "decaying" as they are relatively constant. Those resources could remain unchanged for a long time if stored properly. However, we cannot share all resources in an article with conventional formats. For instance, articles sometimes rely on datasets with large volumes that are shared on personal websites of the resources or external services. These datasets could be of various kinds, including, but not limited to, genomic sequences, machine learning datasets, images, text, video, and audio. Other artifacts that are especially common these days are software code, preprocessing scripts, and binary code. Some of these resources are sometimes shared with their entire history (e.g., Github) or with highly reproducible container environments (e.g., Docker image). Regardless, science is more and more dependent on datasets and code that are widely shared.

The Internet is a significant facet of modern science and a crucial research tool to store and share resources in place of libraries and physical storage (Lawrence & Giles, 2000). Thus, network-accessible information is increasingly being cited by journal articles (Duda & Camp, 2008). While scholarly articles on the Internet are referenced, artifacts used or created during research like software, ontologies, and datasets are also referenced (Klein et al., 2014). Yet resources from the Internet can be unreliable due to the dynamic nature of the web (Duda & Camp, 2008). Multiple researchers have found a consistent lack of persistence in these resources (Klein & Nelson, 2010; Koehler, 1999, 2004). Thus, utilizing storage systems and structures that are fragile for science could cause significant issues for maintaining the value of scientific knowledge.

Predictably, many researchers have found that Internet-stored resources cited by scientific articles suffer from "reference rot" or "link rot" (Klein et al., 2014; Van de Sompel et al., 2014). More specifically, the resource identified by a URI ceases to exist and becomes unavailable. A related phenomenon is "content drift," where resources shared in URIs change over time and cease to be relevant for what was initially referenced. The field of biomedical science has been a pioneer in attempting to avoid these issues but still has some problems in software (Mangul et al., 2018), protocols (Open Science Collaboration, 2015), and datasets (Bonàs-Guarch et al., 2018). Due to the importance and cost of biomedical science, the fact that these resources disappear should be of great concern. We might expect that other fields considered less crucial suffer similar or worse problems. Therefore, we need mechanisms to keep track of this disappearance, drifting, and irrelevance.

Government and funding agencies have been pushing to establish higher standards for data sharing. Around 2003, the National Institutes of Health (NIH) published a policy requiring applications for grants greater than $500,000 to include data-sharing plans (National Institutes of Health, 2003). The National Science Foundation has developed similar policies encouraging data sharing (National Science Foundation, 2011). Countless other institutions recognize the importance of such practice (e.g., see (Milham et al., 2018; Van Horn & Gazzaniga, 2013)) and its impact on science (e.g., see Bonàs-Guarch et al., 2018). Sharing of resources is essential, and we need to better understand the process of its decay.

Why predicting resource longevity might be helpful. Ideally, we would like to predict how resources decay to optimally allocate

when to perform maintenance or simply decide that a resource is no longer relevant. Many systems work based on this principle. For example, libraries with long-term storage systems need to understand how storage decays and therefore need to plan accordingly to update storage systems, ventilation, and other factors (El-Bakry & Mohammed, 2009). Furthermore, in the digital world, long-term storage requires engineers to plan backups, upgrades, and removal policies to ensure a good tradeoff between storing relevant information accurately and removing information that is not used.

In science, we do not think about these issues directly. However, we do share our findings as soon as possible, indirectly implying that resources are most useful sooner rather than later. For example, when we cite previous research, it is standard practice to cite relevant publications recently published (Penders, 2018). The understanding is that research published too long in the past may not translate well to today's research questions (Plavén-Sigray et al., 2017). When publications share software, this is even more apparent. In machine learning, for example, it is common practice that the code and data are shared alongside the software and data storage versions so that other scientists in the future can reproduce the results (Kop, 2020). While research has shown that machine learning might lack reproducibility (Haibe-Kains et al., 2020), machine learning is a field where sharing artifacts beyond publications is standard practice at least. Therefore, scientists, either implicitly or explicitly, are always thinking about the relevance of resources.

Materials and methods

Data

Dataset of URLs in Open Access publications. In this research, we enriched data from (Zeng et al., 2019) that collected the URL links within PubMed Open Access Subset (PMOAS) publications and checked whether they were accessible. In the original dataset of that research, each record consists of fields such as PubMed ID (PMID), URL and URL availability. The study extracted several features from the authors, publications, and the URLs contained therein, including domain country, size of the resource, number of affiliations, *h*-index of the journal, age of the URL, number of references, length of the abstract, and the frequency of use of URL in other publications. The study found that the frequency of reuses and the "int" (international), "org" (mostly non-profit organizations), and "gov" (governmental) domains are most predictive of a link being alive. Conversely, it found that the length of the URL's path (after the domain), and the Indian, European Union, and Chinese domains were most predictive of resources not being accessible. However, that study did not investigate how long a project would last and did not consider important information related to citations, the technology used, and the prestige of the authors.

Dataset enrichment. Here we aimed to predict the longevity of resources based on the features present when authors first share the resources. The features we examined were not just related to the link itself, but the technology used in the linked resources and other factors from the publication (e.g., the authors, and venue). We also used the Microsoft Academic Graph (MAG), a large-scale knowledge graph consisting of six types of entities (i.e., paper, author, institution, venue, event and field of study) and the relationship between them (Wang et al., 2019). To associate the dataset from (Zeng et al., 2019) with MAG, we first searched the PMID in PubMed Central APIs to collect its digital object identifier (DOI), we then matched the DOI in MAG to find the corresponding paper. Further, we extracted article-related features and author-related features from MAG (as shown in Table 1). The reconciled relationships helped predict the longevity of resources shared in a scientific publication.

To reconstruct how web resources (i.e., URLs) looked at the time of publication, we built a parallelized python scraping program to enrich the dataset from (Zeng et al., 2019) In particular, the Python program integrated the Wayback Machine API[1], which allowed us to access the history of a web resource during an arbitrary point in the past.

Building a scraping application was an engineering challenge. Thus, we broke the program into four components: a global shared data cache, a database access component, an outbound traffic rate limiting component, and an execution pool to run tasks in threads simultaneously. The execution pool that receives requests to create jobs of various kinds, simplified the program parallelization. First, the database access component reads a list of candidate URLs from a file, creates a job to initialize a database of records scraping, and loads all the pending records into a global shared cache. After that, the application schedules another two jobs in parallel to download the web page and clean the browser cache the Python library used periodically. The Wayback Machine API restricts the call to 15 per minute for an application. To cope with this rate limit policy, we implemented a token bucket algorithm as a rate-limiting component to control the outbound traffic. Since less than 10% of the web resources were unavailable when building the dataset, we set the capacity of this token bucket rate limiter to 500, which was the same as the size of the execution pool.

The scraping program attempts to communicate with the original URL and uses the Wayback Machine as its complementary resource (Figs. 1 and 2). At the beginning of the scraping, we filtered out several types of URLs. First, as we were analyzing the quality of the website, we ruled out URLs with FTP protocol which represent resources in a file system. We also excluded URLs that were unavailable in the original destination or not archived in the Wayback Machine. We logged the last available date for all of the rest URLs. If we scraped the URL from its original location, the current timestamp would be recorded as its last known available time. Otherwise, we filled it with the archive's timestamp representing the last time it appeared in the Wayback Machine. Finally, we extracted information from the HTML's source code and dependencies. In particular, we extracted features of the technical structure, such as the presence of the iframe, the type of the suffix and charset, and the number of internal and external JavaScript dependencies.

We noticed that a large portion of our URLs were accessible at the time of the study; about 90% of the URLs were alive when we fetched them in April of 2020. Their lifespan could be more than the age calculated by subtracting it from the year it first becomes available. In addition, considering the HTTP IO is an expensive operation, it took us more than a month to finish the scraping for the dataset from (Zeng et al., 2019). This implies that timestamps for those available URLs were not captured at the same time. For resources only available on Wayback Machine, they would have a higher likelihood of being crawled if they were well linked. In general, the strategy Wayback Machine scrapes for web resources affected the data distribution of the last timestamp of the URL's snapshot.

Methods

Feature engineering. Considering that we were trying to predict the longevity of a resource shared in a scientific article, in this section we explain the features we extracted from the URL itself, features from the website's HTML and resources, and features extracted from MAG about authors of the publication, the

Table 1 Features and their rationale. Some of the features were inherited from (Zeng et al., 2019).

Feature group	Feature	Description	Purpose of the feature
URL information	1. Protocol type	It's the type of the protocol stated in the URL, it can be HTTP or HTTPS. Other protocols like FTP are not included in the analysis.	Since the lifespan of the SSL resource needs a renewal every other 27 months, a web resource with HTTPS protocol will need more efforts to keep it available. An organization may plan the maintenance ahead. Otherwise, the research result will disappear along with the expiration of SSL due to the lack of project plan and maintenance.
	2. Depth of the path in a URI	The URI depth is related to the design of the web page. For example, a web page with a REST architecture style uses the path hierarchy to identify the location of a resource.	The depth of the URI represents a hierarchical structure in a web domain. Eg, the REST API is a well-designed URL pattern that publishes the resources of an application to the outside in an organized manner. An easy-to-remember URI is beneficial to the lifespan of a web resource.
	3. www or non-www URI	Www is one of the popular subdomains of a website. From the user's perspective, it is the synonym of the front page which content is bonded to the primary domain. The website developer usually routes the www subdomain to its primary domain by sending a redirect HTTP request. The web resource may not be available without proper maintenance.	From the perspective of SEO, a domain with or without www will have no difference being included in the search engine. However, to unify the content of these 2 domains requires extra effort, because the www suffix is the subdomain of the main domain. The website engineer needs to redirect (HTTP code 300 family) from one to another. This feature will help us to figure out the relationship between the maintenance of the website and its lifespan.
	4. The level of the subdomain	Self-explanatory feature	
	5. The number of query parameter in the URI	How many parameters are included in the HTTP Get request.	The HTTP Get request accepts additional query parameters in the URL. The argument list depends on the design of the backend system. The lack of the sustainability of the system design is detrimental to the availability of a web resource, as referred links in a paper cannot be revised after being published.
	6. Domain suffix with org 7. Domain suffix with int		From (Zeng et al., 2019), the org and int suffix are said to be one of the important factors that positively correlated to the lifespan of the website.
	8. Domain suffix with jp		According to the research, the top non-US tools from Japan maintain a high reputation and are frequently cited. It is worthy to explore the suffix with JP to see how its impact on the lifespan of the web resources.
	9. Domain suffix with gov		As above
	10. Indian domain (.in)		As stated in (Zeng et al., 2019),
	11. Domain suffix with cn 12. Domain suffix with eu 13. Domain suffix with kr		Resources with in/cn/eu/kr negatively contribute to the longevity of a website.
	14. Is the web resource explicitly accessed through port	A flag to identify the existence of the port information in the web resource.	By default, a direct access to a website implicitly connects to port 80. A URI with explicit port access may not be a user-friendly design and could be revised in a later version.
HTML contents and technology	15. The size of the HTML source code (in string length)	Self-explanatory feature	The size of the HTML code is one of the analytical factors in our research.
	16. The length of the HTML title in the source code	Self-explanatory feature	The HTML title abstracts the details of the web resource, which increases the readability of the website.

Table 1 (continued)

Feature group	Feature	Description	Purpose of the feature
	17. The number of internal JavaScript file	Self-explanatory feature	While hosting the JavaScript file inside a domain would increase the complexity of the maintenance process, it may give the full autonomy on extension development of the website. We included this feature because we want to explore how this property affects the lifespan of the web resources.
	18. The number of external JavaScript file	A JavaScript library imported from the location hosted by the same web domain	A JavaScript library imported from a location outside the web domain. Whether this exterior JavaScript resource is accessible or not, its availability relies on the external resource itself.
	19. The charset specified in the HTML source code	A categorical variable that defines the type of the char encoding specified in the web page.	A charset classifier controls the encoding of the page. The content of the web page is vulnerable to be messed up if the decoding charset in the browser is inconsistent with the one specified in the web page. Since the UTF-8 encoding is widely used across many kinds of websites, we presume using UTF-8 encoding would help to extend the lifespan of the web resource.
	20. HTML or HTML5	An indicator to distinguish the HTML5 page from the regular HTML.	The first HTML5 opened to the public was created on 22 January 2008. So the HTML5 indicator can be considered as a representation of chronology information. While the new features of HTML5 introduced the ability of presenting a variety of information to the webpage, processing the data with heterogeneous structure in the webpage requires more effort to be utilized by the third party.
	21. The iFrame tag usage indicator	Whether the web page use the iFrame tag or not	The iFrame tag can be used to control the layout of the page, or to address the Cross-origin resource sharing issue. The single-page application design is extensively used in the modern website. If the source code of a website utilized the iframe, then this page could be built based on the technology from the previous generation. So this feature is a chronological indicator.
	22. The number of **hyperlinks** on the web page	The number of web links that represent other resources on the internet.	Publishing the information through the web link is a complementary way of sharing the knowledge. Extensively using web resources or conventional text conducted article sharing are different ways of researching. Analyzing this feature helps to identify how the way of page organizing impacts the longevity of the web resource.
Article references	23. Total number of publications referenced by the paper	Self-explanatory feature total_num_of_paper_referencing	The paper consolidates the information of other cited papers. The referred material is the extension of the paper. This feature helps to measure how the paper referral impacts its lifespan.
	24. Total number of authors referenced by the paper	Self-explanatory feature num_of_author_referncing	Researchers have their own study field. Although we cannot claim the paper is interdisciplinary research by solely relying on the number of the author, the number does tell us the degree of the innovation of the paper.

Table 1 (continued)

Feature group	Feature	Description	Purpose of the feature
	25. Total number of affiliations referenced by the paper	Self-explanatory feature num_of_affiliation_referencing	The study area among researchers in the same affiliation is closer than that in the different affiliation. Collaborating this feature with the total number of referred affiliation in the analysis helps to explore how the concentration of the research field contributes to the longevity of resources.
	26. Total number of journals referenced by the paper	It measures the diversity(number) of the different referred journals. num_of_journal_referencing	
	27. The total number of self-citations in a paper.	The number of cited papers belongs to the authors themselves total_num_of_author_self_citation	This feature evaluates how strict the author of this paper is on the previous research direction.
	28. The number of citations of a paper that is published by the same affiliations.	Self-explained feature. total_num_of_affiliation_self_citation	The communication among researching is more feasible to carry out in the same affiliation. But it is vulnerable to get a homogeneous result if the paper refers to more resources from the same affiliation.
	29. The number of citations of a paper that is published **in** the same journals.	Self-explained feature. total_num_of_journal_self_citation	Each journal has its unique criteria of evaluating and accepting the paper. It reflects the research quality of the paper. Gauging this number helps to measure the impacts of diversity of the quality to the lifespan of web resources.
	30. The average year of publications referenced by the paper	Self-explained feature. avg_year	A chronology information that represents the overall years of the cited paper.
	31. The earliest year of publications referenced by the paper	Self-explained feature min_year	See feature 30
	32. The latest year of publications referenced by the paper	Self-explained feature max_year	See feature 30
	33. The median year of publications cited by the paper	Self-explained feature median	See feature 30
Authors quantity	34. The number of authors	How many authors collaborated on this paper. num_of_author	Behind this number is the effort spent by each author. It indirectly reflects the scale of the paper in terms of the authors.
	35. The number of affiliations that published the paper's citations.	How many affiliations are engaged in the citations. paper_unique_affiliation	
	36. The total number of authors of citations of a paper.	How many authors (with duplication) wrote the citations of a paper. total_num_of_author_citing,	Behind this number is the effort spent by each author. It indirectly reflects the scale of the paper in terms of the resources it used.
	37. The total number of journals that published the paper's citations.	total_num_of_journal_citing	
	38. The total number of affiliations that published the citations of the paper	How may affiliations (with duplication) were involved in this particular paper total_num_of_affiliation_referencing	This feature measures the degree of the collaboration among affiliations.
Authors' prestige/ seniority/impact	39. The arithmetic average level of H-Index	It is the number averaged by all years of h-index among all the authors of a paper. avg_hindex	These features use the H-index of the authors to estimate the longevity of resources.
	40. The level of H-Index of the primary author	Self-explained feature first_author_hindex	
	41. The level of H-Index of the last author	Self-explained feature last_author_hindex	
	42. The level of H-Index of authors other than the first and last.	Self-explained feature avg_mid_author_hindex	This feature evaluates the academia degree of the minority contributors of this paper.

Table 1 (continued)

Feature group	Feature	Description	Purpose of the feature
Dependent variable	43. The last available year of the website	The last available day of the web resource	We scaled the dependent variable by using the number 1990 to subtract the year it was last available, because the first web page of the world was published in 1991.

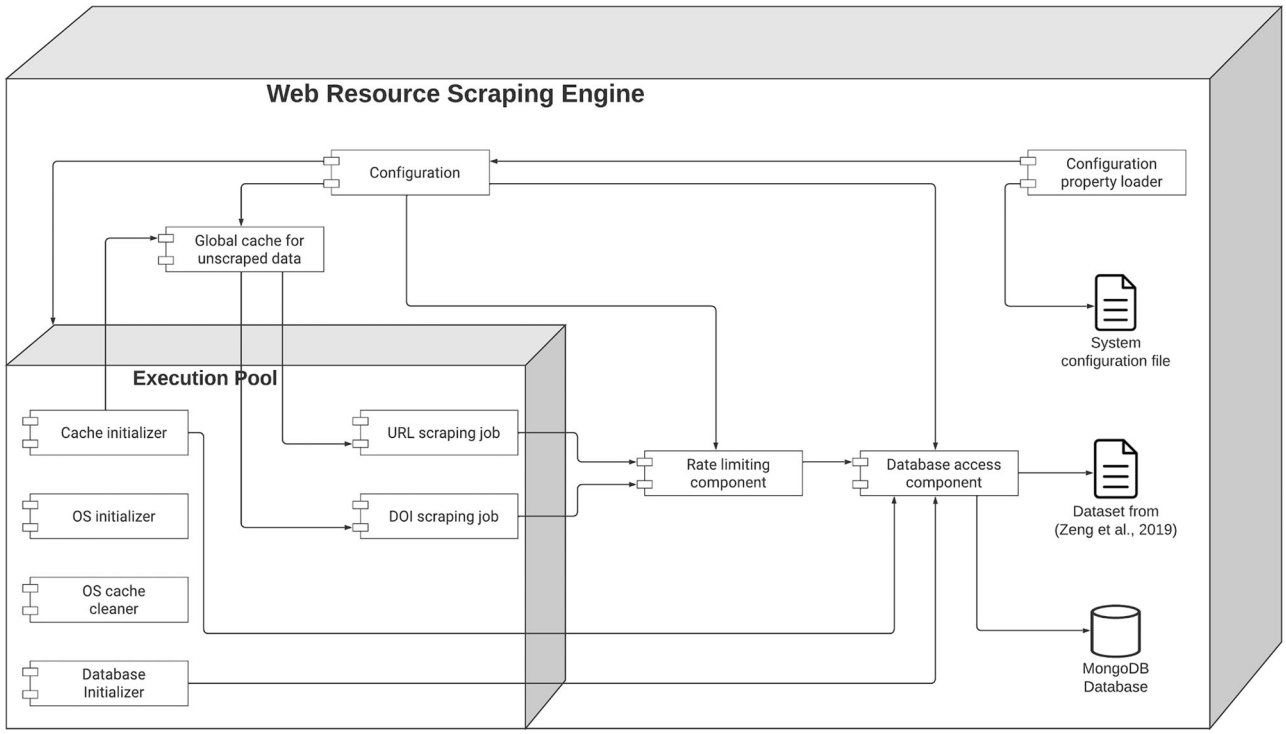

Fig. 1 The architecture of the scraping program.

institution, the journal, and the citation structure of the article itself. This feature engineering resulted in 42 independent variables and 1 dependent variable (Table 1).

Models. We used two types of models for our modeling and prediction. The first model focuses on interpretability and has desirable statistical properties based on a censored regression model. The other model is based on Random Forest.

Censored model of resource decay. Most of the URLs we analyze are still available, and therefore, we do not know their remaining lifespan. However, we wanted to use this information to inform the model that predicts longevity. To achieve this, we used techniques from censored regression models. These types of regression models split the prediction into two parts: one predicts the availability of the resource, and the other predicts its unobservable remaining lifespan.

In particular, we inherited and advanced the Tobit model to perform a censored regression. The following paragraphs will explain in detail how such a model fit works using a Bayesian maximum a posteriori framework. To mitigate overfitting, our approach imposes a prior distribution over the parameters and maximizes the posterior distribution of the parameters rather than simply the likelihood function. Our advanced Tobit model constructs a data likelihood $L(\beta, \sigma)$ based on the parameters β and σ and weighs that by the prior distribution $P(\beta)$ over the parameters. Concretely, the likelihood function is given by

$$L(\beta,\sigma) = \prod \left[\Phi\left(\frac{a - X^T\beta}{\sigma}\right)\right]^{I^a} \times \prod \left[\frac{\varphi\left(\frac{y - X^T\beta}{s}\right)}{\sigma}\right]^{(1-I^a-I^b)} \times \prod \left[\Phi\left(\frac{X^T\beta - b}{\sigma}\right)\right]^{I^b}$$

and we introduce an assumption that the prior distribution follows a mixture of Laplacian distribution and Gaussian distribution, which is similar to assuming L2 and L1 regularization parameters as in Elastic Net regularization:

$$P(\beta) = \left(f_{\text{Gaussian}}(\beta)^{(1-\alpha)} \cdot f_{\text{Laplacian}}(\beta)^{\alpha}\right)^{\lambda}$$

Then the posterior distribution is given by

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

$$P(\beta|Y) = \frac{P(Y|\beta) \cdot P(\beta)}{P(Y)}$$

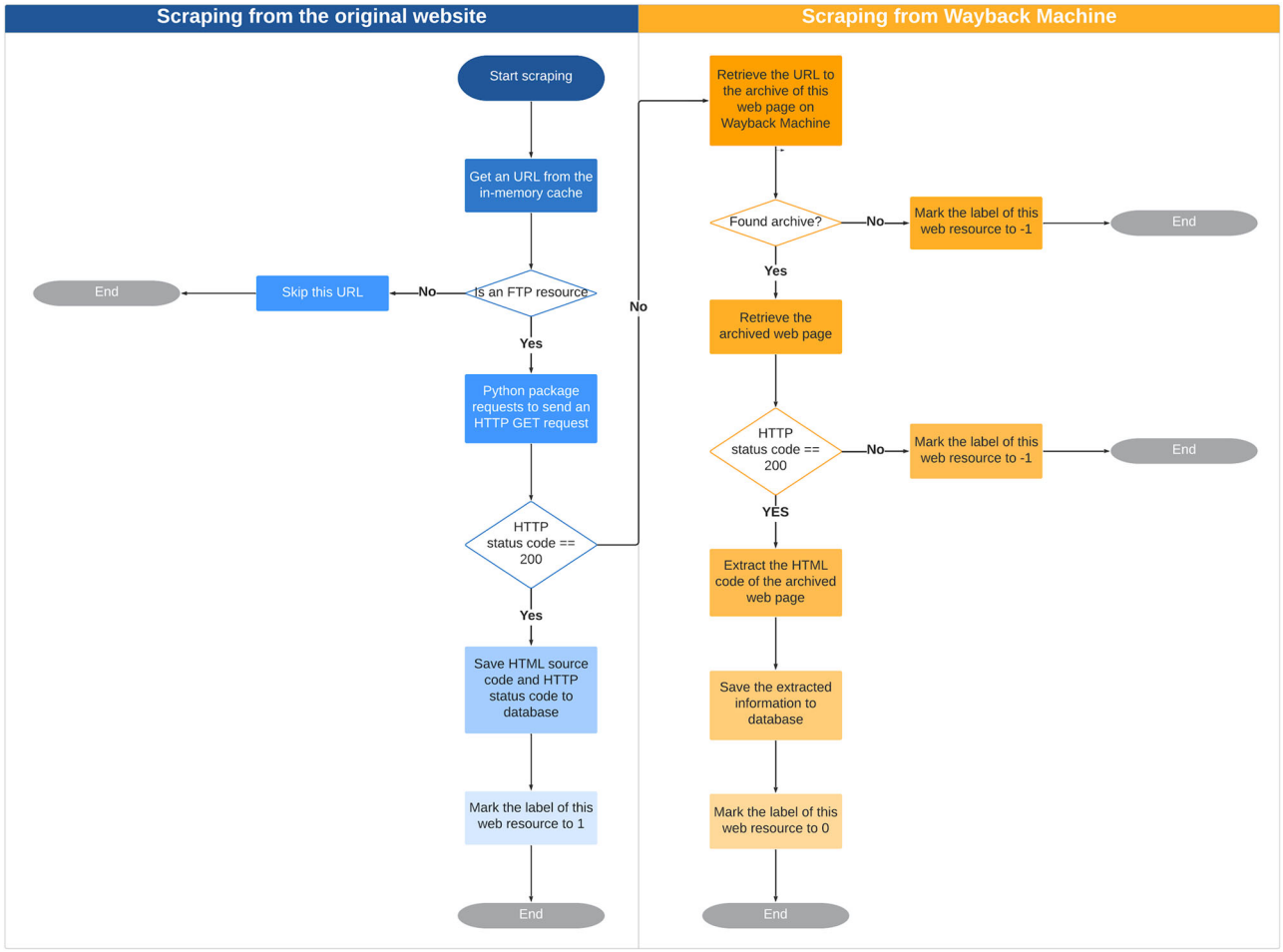

Fig. 2 The workflow chart of the scraping program.

$$= \frac{L(\beta,\sigma)\cdot P(\beta)}{P(Y)}$$

$$P(\beta|Y) = \frac{L(\beta,\sigma)\cdot\left(f_{\text{Gaussian}}(\beta)^{(1-\alpha)}\cdot f_{\text{Laplacian}}(\beta)^{\alpha}\right)^{\lambda}}{P(Y)}$$

The log-posterior distribution will then take the more favorable optimization form

$$\begin{aligned}\log(P(\beta|Y)) &= \log\left(\frac{L(\beta,\sigma)\cdot f_{\text{Gaussian}}(\beta)^{(1-\alpha)}\cdot f_{\text{Laplacian}}(\beta)^{\alpha}}{P(Y)}\right)\\ \log(P(\beta|Y)) &= \log(L(\beta,\sigma))\\ &+ \log\left(f_{\text{Gaussian}}(\beta)^{(1-\alpha)\cdot\lambda}\right)\\ &+ \log\left(f_{\text{Gaussian}}(\beta)^{(1-\alpha)\cdot\lambda}\right)\\ &- \log(P(y))\\ &= \log(L(\beta,\sigma))\\ &+\lambda\cdot(1-\alpha)\cdot\log(f_{\text{Gaussian}}(\beta))\\ &+\lambda\cdot\alpha\cdot\log\left(f_{\text{Laplacian}}(\beta)\right)\\ &-\log(P(Y))\end{aligned}$$

As we estimate the probability in terms of the β, we can eliminate the constant term evidence and apply MLE to it.

$$\begin{aligned}\hat{\beta} &= \arg\max P(\beta|Y)\\ &= \arg\max\log(P(\beta|Y))\\ &= \arg\max\log(L(\beta|Y))\\ &\quad+\lambda\cdot(1-\alpha)\cdot\log(f_{\text{Gaussian}}(\beta))\\ &\quad+\lambda\cdot\alpha\cdot\log\left(f_{\text{Laplacian}}(\beta)\right)\\ &\quad-\log(P(Y))\\ &= \arg\max\log(L(\beta,\sigma))\\ &\quad+\lambda\cdot(1-\alpha)\cdot\log(f_{\text{Gaussian}}(\beta))\\ &\quad+\lambda\cdot\alpha\cdot\log\left(f_{\text{Laplacian}}(\beta)\right)\end{aligned}$$

We **then** use an L-BFGS-B optimization which requires to provide the gradient, which will simply become

$$\frac{\Delta\log P(\beta|\sigma)}{\Delta\beta} = \frac{\Delta\log(L(\beta,\sigma))+\lambda\cdot(1-\alpha)\cdot\log(f_{\text{Gaussian}}(\beta))+\lambda\cdot\alpha\cdot\log\left(f_{\text{Laplacian}}(\beta)\right)}{\Delta\beta}$$

$$\frac{\Delta \log P(\beta|\sigma)}{\Delta\beta} = \frac{\Delta \log(L(\beta,\sigma))}{\Delta\beta} + \frac{\lambda\cdot(1-\alpha)\cdot\log(f_{\text{Gaussian}}(\beta))}{\Delta\beta} + \frac{\lambda\cdot\alpha\cdot\log\left(f_{\text{Laplacian}}(\beta)\right)}{\Delta\beta}$$

For a Gaussian distribution:

$$\begin{aligned}\frac{\log(f_{\text{Gaussian}}(x))}{\Delta x} &= \frac{\log\left(\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)}{\Delta x}\\ &= \frac{\log\left(\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\right)+\log\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)}{\Delta x}\\ &= \frac{\log\left(\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\right)}{\Delta x} + \frac{\log\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)}{\Delta x}\\ &= \frac{\log\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)}{\Delta x}\\ &= \frac{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}{\Delta x}\\ &= -\frac{x}{\sigma}\end{aligned}$$

We assume that the standard deviation in the prior is a unit value, further simplifying it into

$$\frac{\log(f_{\text{Gaussian}}(x))}{\Delta x} = -x$$

Similarly, for the Laplacian distribution

$$\begin{aligned}\frac{\log\left(f_{\text{Laplacian}}(x)\right)}{\Delta x} &= \frac{\log\left(\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}\right)}{\Delta x}\\ &= \frac{\log\left(\frac{1}{2b}\right)+\log\left(e^{-\frac{|x-\mu|}{b}}\right)}{\Delta x}\\ &= \frac{\log\left(\frac{1}{2b}\right)}{\Delta x} + \frac{\log\left(e^{-\frac{|x-\mu|}{b}}\right)}{\Delta x}\\ &= \frac{\log\left(e^{-\frac{|x-\mu|}{b}}\right)}{\Delta x}\\ &= \frac{-\frac{|x-\mu|}{b}}{\Delta x}\end{aligned}$$

By applying the subgradient, we get the expression below

$$\frac{\log\left(f_{\text{Laplacian}}(x)\right)}{\Delta x} = -1 \text{ if } x>0$$

$$\frac{\log\left(f_{\text{Laplacian}}(x)\right)}{\Delta x} = 0 \text{ if } x=0$$

$$\frac{\log\left(f_{\text{Laplacian}}(x)\right)}{\Delta x} = 1 \text{ if } x<0$$

Let the functions $L1(\beta)$ and $L2(\beta)$ be

$$L1(\beta) = \{-1 \text{ if } x_i>0;\ 0 \text{ if } x_i=0;\ 1 \text{ if } x_i<0\}$$

$$L2(\beta) = X$$

we can express the gradient of log a posteriori as

$$\begin{aligned}\frac{\Delta \log(L(\beta,\sigma))}{\Delta\beta} = \sum_{i=1}^{N}\Bigg[&-I^a\frac{\varphi\left(\frac{a-X^T\beta}{\sigma}\right)}{\Phi\left(\frac{a-X^T\beta}{\sigma}\right)}\frac{x}{\sigma} + I^b\frac{\varphi\left(\frac{X^T\beta-b}{\sigma}\right)}{\Phi\left(\frac{X^T\beta-b}{\sigma}\right)}\frac{X}{\sigma}\\ &+\left(1-I^a-I^b\right)\frac{Y-X^T\beta}{\sigma}\frac{X}{\sigma}\\ &+\lambda\cdot\left(\alpha\cdot L1(\beta)+(1-\alpha\cdot)L2(\beta)\right)\Bigg]\end{aligned}$$

$$\frac{\Delta \log(L(\beta,\sigma))}{\Delta \log(\sigma)} = \sum_{i=1}^{N}\Bigg[-I^a\frac{\varphi\left(\frac{a-X^T\beta}{\sigma}\right)}{\Phi\left(\frac{a-X^T\beta}{\sigma}\right)}\frac{a-X^T\beta}{\sigma} + I^b\frac{\varphi\left(\frac{X^T\beta-b}{\sigma}\right)}{\Phi\left(\frac{X^T\beta-b}{\sigma}\right)}\frac{X^T\beta-b}{\sigma} + \left(1-I^a-I^b\right)\left(\left(\frac{Y-X^T\beta}{\sigma}\right)^2-1\right)\Bigg]$$

We use the above gradients together with the Python package `scipy.optimize` to find the optimal parameters. Because we needed to estimate the uncertainty about these parameters, we use bootstrapping estimation of the posterior distribution (Murphy, 2012).

The Tobit model mentioned above processes an important property of being easily understandable through its $\hat{\beta}$ features: they are a simple regression that will linearly influence how long a resource is believed to last. We can take the weight of a feature j (see feature engineering section above), and we can try to understand how a unit increase in such as feature produces a change of $\hat{\beta}_j$ units to the output Y. However, the problem with this model is that it considers features to influence the outcome linearly and does not consider interactions between features unless we explicitly add them to the model.

Random Forest model. Random Forest aims at solving some of the shortcomings of linear models. Particularly, they are based on a combination of classifiers or regressions known as decision trees, and therefore produce non-linear relationships between the features and outcome. However, because decision trees tend to overfit the data, Random Forest bootstraps the data to reduce the model's variance. A further improvement that Random Forest applies is to randomly sample the features themselves, essentially making each of the trees substantially different from each other compared to simple bootstrapping. The effect of such *bagging* (i.e., bootstrapped aggregation) and feature bootstrapping is that the variance decreases with more trees. Random Forests have been shown to work remarkably well without much hyperparameter tuning (James et al., 2013).

Model interpretability. Even though the Tobit model is highly interpretable, it is challenging to compare the weights of that model to the feature importance from the Random Forest. We use SHAP values to compare the feature importances in our models. SHAP importances can fill this gap. Essentially, SHAP values estimate how much a feature makes a classifier or regression change its prediction compared, on average, to all possible combinations of their features (Lundberg et al., 2020). Importantly, we can apply this definition to any model. Therefore, we can provide some clues as to which features are more critical in our predictions, whether using the Tobit model or Random Forest.

Results

Prediction performance. In our dataset, we have 58,871 resources shared as URLs within scientific publications. The average lifespan of these resources is 19.45 years (Fig. 3). Because we have the entire survival history of each URL, we can attempt to reproduce the predictability of survival presented in (Zeng et al., 2019). Indeed, by using the features presented in our feature engineering section (see Methods), we estimated a result similar to the conclusion in the previous study: the URL domain is one of the factors that determine the longevity of a web resource and the predictability of a resource being unavailable in (Zeng et al., 2019).

To better understand these features, we used to train the model, we studied the simple correlations between the features proposed in methods for both resources that are available today and those that are not. We found that for both the unavailable and available resources (Figs. 4 and 5, respectively), the features

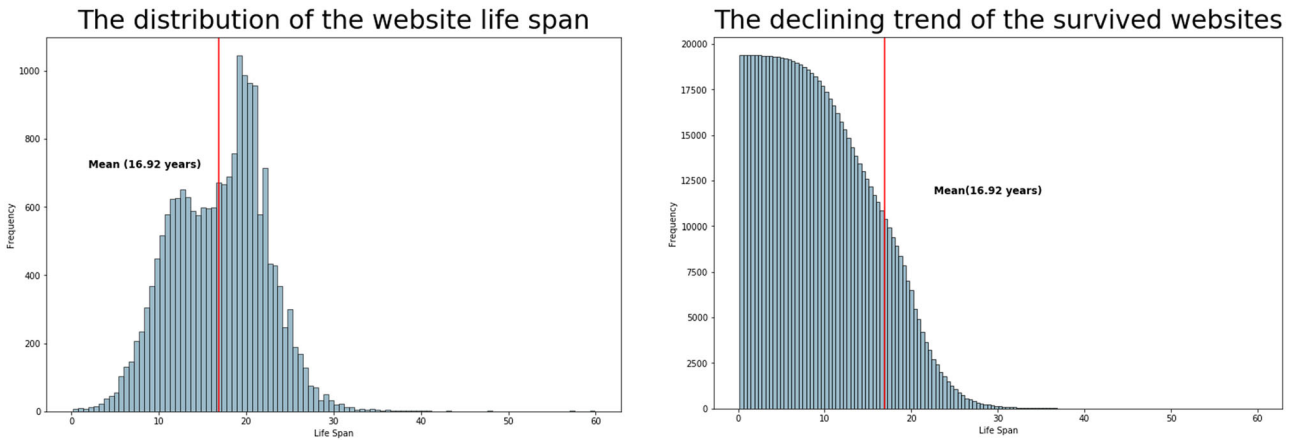

Fig. 3 Life span distribution of our dataset. Left panel: distribution of life span. Right panel: the probability of a resource having more than a given life span. The mean life span is 19.45 years.

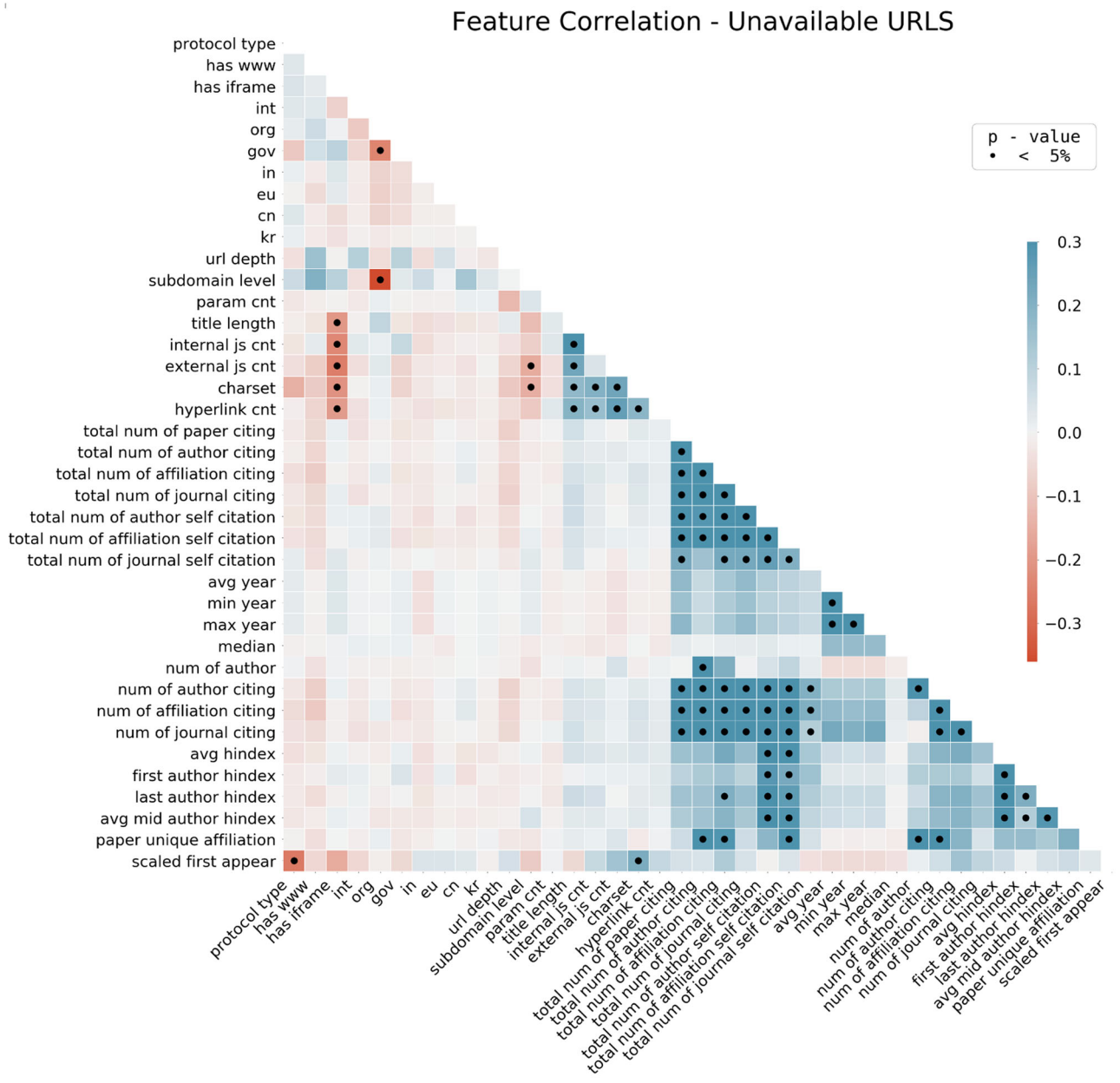

Fig. 4 Correlations between features and their significance values for unavailable resources. Naturally, there is a heavy positive and significant correlation among features related to citations and references.

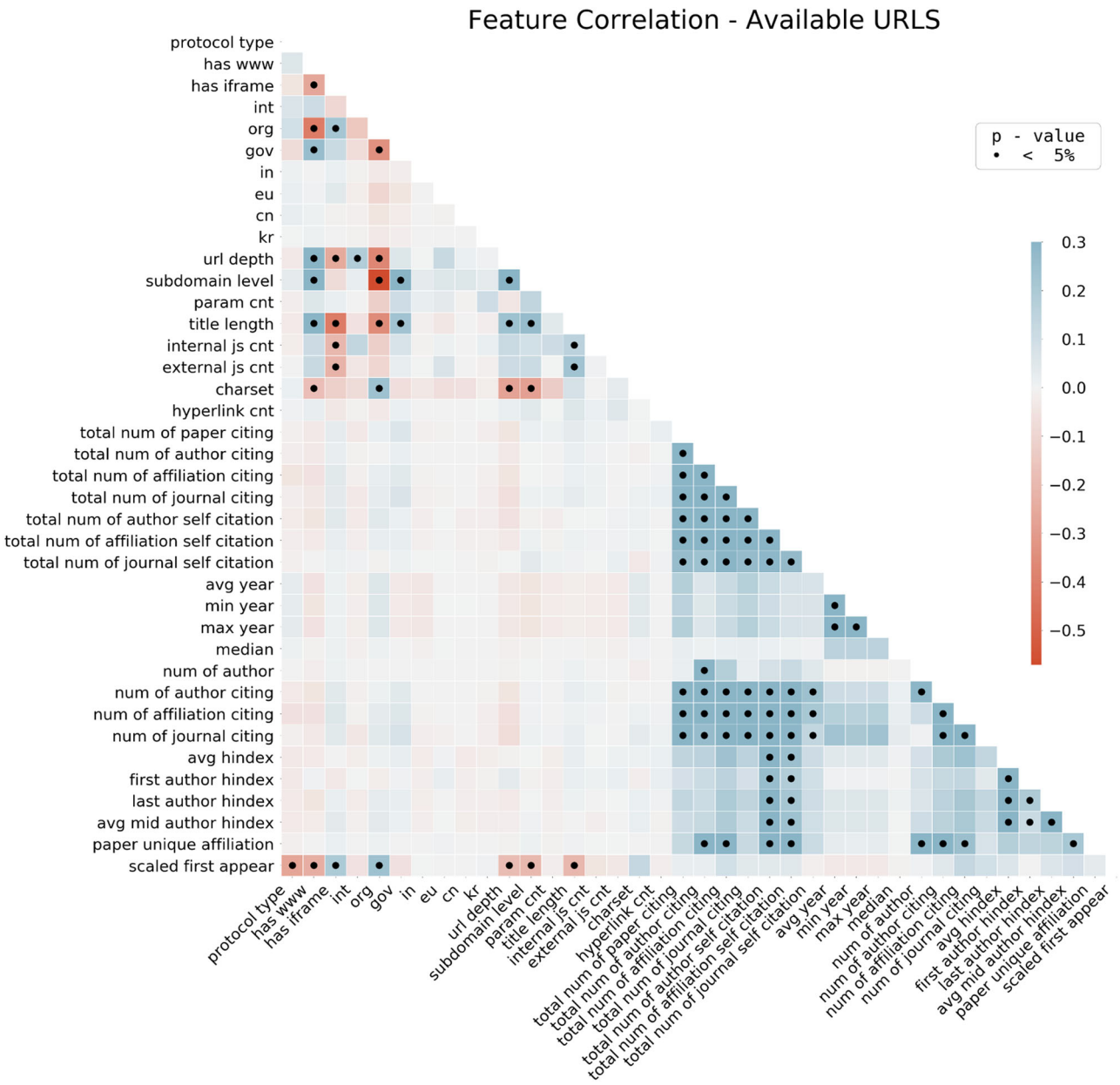

Fig. 5 Correlations between features and their significance values for available resources. Similar patterns compared to unavailable resources (Fig. 4), but with lower correlation values.

corresponding to citations are positively correlated as expected. However, the features were less correlated in the available resources (Fig. 5) correlations, suggesting that they were more diverse.

To understand how each factor contributes to the longevity of a web resource, we experimented with an elastic net regularized linear model for our dataset, which resulted in an estimated validation R squared of 0.164. However, our data was censored (i.e., the remaining lifespan of those web resources is unobservable), so we proceeded with a more sophisticated Tobit model presented in the "Methods" section.

In our dataset, a high percentage of the resources were still available at the time of the study (90.49%). However, we had to account for those for which we did not know when they disappeared (9.51%). In this case, we could not compute the performance using the traditional formulation of R squared, but instead, we had to do it with pseudo-R-square

$$\text{pseudo } R^2 = 1 - \frac{\log p(Y|\beta)}{\log p(Y)},$$

where $p(Y|\beta)$ comes from the Tobit model. With this model, we estimated our pseudo-R-squared to be 0.045 in cross-validation. The pseudo R squared and the R squared could not be compared because the units are much different. Finally, pseudo R squared parameters based on log likelihood tend to be much lower than traditional R squared because log likelihoods can be much bigger numbers compared to traditional variance (Mbachu et al., 2012).

How a resource is shared but not who shares it is important for longevity. We also analyzed what estimators in the Tobit model are important to the prediction. To estimate the uncertainty

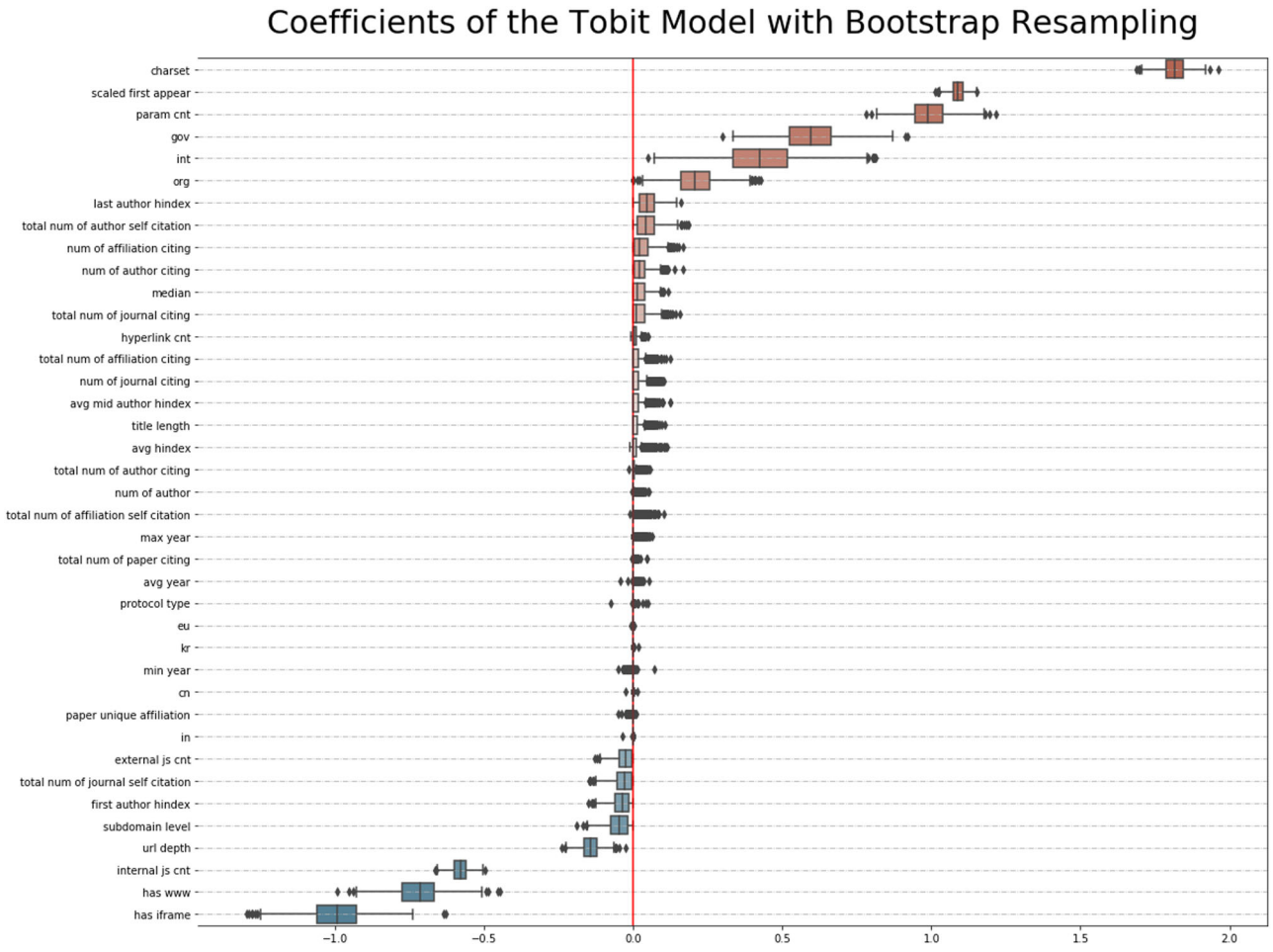

Fig. 6 Feature importance for predicting the longevity of resources shared by scientific articles.

about each of the models, we bootstrapped the parameters as well (see Methods)[2]. Because we standardized the features, we could, in principle, compare them one to another (Fig. 4). Using this idea, we conclude that features related to technology (charset, parameter count, and domain for positive factors; a locally hosted javascript library represented by feature `internal_js_cnt` and whether it has an iFrame) are contributing factors to the prediction. On the other hand, we did not find strong predictability of other factors such as the *h* index, number of citations and references, and the journal's prestige. These results suggest that technology plays a significant role in making sure that resources stay alive longer than average.

One of the issues with the Tobit model we proposed is that it does not capture potential non-linear relationships between the features and the longevity of the data. To account for this, we also use a Random Forest model and compare the features that it found relevant to those found significantly different from zero in the Tobit model. Indeed, the coefficients that the Random Forest thinks are most important are highly related to the Tobit model (Fig. 7), including the use of an old Javascript library (`internal_js_cnt feature`) and whether it uses modern HTML structure (`charset` feature).

One of the issues with feature importance in Random Forest is that it does not inform us in which direction the prediction is happening. We use the SHAP values to do this additional analysis (see Methods, Fig. 8, right panel). This analysis, in principle, allows us to compare both models and see if they make similar predictions. Indeed, we find that the SHAP value of the Tobit model is **similarly close** (Fig. 8, left panel). In both, technical-related features are still at the top of the feature list. Unlike the Tobit model, the SHAP values of many features in Random Forest are mixed with the lower value. Only the charset and internal_js_cnt have a polarity effect over the average prediction. The feature scaled_first_appear is the top contributor pushing up the average prediction, followed by three features max_year, `internal_js_cnt`, and title_length. Like the SHAP value in the Tobit model, the technical-related features contribute most to the average prediction.

Discussion

This research aims to understand the factors that predict the longevity of resources shared in scientific publications. First, we used a large sample of open access publications to obtain the URLs of such resources. Then, we reconstructed how these resources looked throughout time by using archive.org's Wayback Machine. This analysis allowed us to understand the factors of the technology used when the authors shared the resource (e.g., type of HTML or Javascript used). Finally, we crosslinked the publication information with information related to the author's references, affiliation, and journal's prestige. Overall, we found that the factors relevant in the prediction were mostly related to the technology used to share the resources more than prestige. We now discuss our findings of our research, limitations, and future work.

The feature importance analysis provides interesting insights into factors related to longevity (Figs. 6–8). First, the lifespan of a resource is predicted to be shorter if the article that shares it has

Fig. 7 Feature importance from Random Forest. The essential features are mostly related to the technology used for sharing data.

many self-citations. While self-citations are not a negative factor per se (Kulkarni et al., 2011), it is a reasonable finding as unusually self-cited work might not be used as much by others. Therefore, there is no motivation to keep it up-to-date and available. Second, the count of hyperlinks in the HTML and the depth of the URL behaves quite identically. A longer URL will likely have a more complex structure, and we would expect it to be harder to maintain. Also, we found that the iframe tag is negatively associated with longevity. This trend is reasonable because the iframe tag is considered an outdated design that is hard to maintain and prone to malicious payloads (Provos et al., 2008). Finally, the categorical feature charset is negatively correlated to the longevity of a web resource.

Even though we said that factors related to prestige were not comparatively as crucial as the technology, it entails helpful information. We found, for example, that the number of different affiliations cited is more critical than the number of other scientists being cited or the prestige (e.g., h-index) of the authors. This trend suggests that the more we cite institutions instead of individual scientists, the more likely we are to care about the resources we are citing—publications with more institutions are cited more often (Gazni & Didegah, 2011), although some have found inter-collaboration is also important (Fu et al., 2018). Such interpretation may influence how we see the maintenance and relevance of resources shared in scientific publications.In the future, we will explore more subtle factors of individuals, affiliations, and countries, including their data management and archiving policy.

Previous research has found some similar trends about resource availability. In the work of (Zeng et al., 2019), the authors found that certain domains such as India and China are more likely to have resources unavailable and that domains such as .org and .gov are more related to resources being available. Contrary to Zeng et al., 2019, we found that the European Union is positively associated with longevity. Nevertheless, we think the first set of factors is related to the trend of technology use in Indian and Chinese higher education institutions, which might be lenient at that moment. Evidence in information systems suggest that citations to internet-based resources are becoming more and more prevalent in both countries (Chen et al., 2014; Milham et al., 2018; Sampath Kumar & Prithvi Raj, 2012). Conversely, the

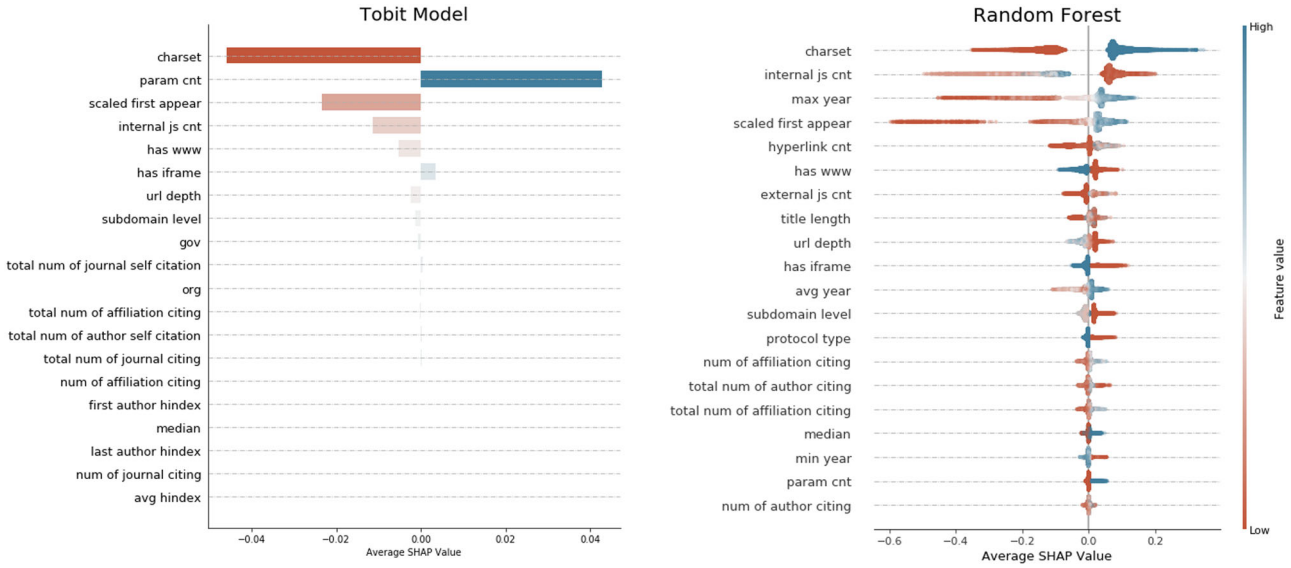

Fig. 8 SHAP feature importance for the Tobit model (left panel) and Random Forest (right panel).

resources shared in .gov domains tend to be governmental institutions, which, comparatively, may have more established standards and resources for data management. Similarly, domains such as .org might have similar better-established governance. Future research will explore data management policy and archiving across countries and types of institutions.

There are several limitations to our research. First, we were analyzing publications in biomedical sciences. Perhaps the technology used was different in more technically sophisticated fields such as engineering and computational science. Also, the factors that affect resources being shared might not be the same. Previous research, for example, showed that scientists are accustomed to sharing data and code through systems such as figshare.com, GitHub and Amazon's AWS S3 cloud storage (Plantin et al., 2018). However, it has also been found that research might not be reproducible given that the resources are available. For example, usage instructions for data might be absent, computer programs in papers might not run, or the results might not match what is reported in the original article (Peng, 2011). It is worth noting that the same could happen in our sample of data. For example, even if we found that resources are "available," we did not know if they are usable. Previous research has found that Excel files shared alongside scientific articles, for example, might contain formulas with errors (Ziemann et al., 2016). Similar problems could be present in biomedical sciences. Future research will explore this subtle but harder-to-examine difference.

We limited our analysis only to scientific journals rather than conferences or other types of events. However, because conferences tend to present results faster (Valcher, 2015), we would expect that their resources would be shared with less priority, ultimately affecting their ability to be maintained and available. Many initiatives were trying to make resource sharing an essential part of knowledge production (Baker & Millerand, 2010; Kaye et al., 2009; Mannheimer et al., 2019). Significantly, the motivation of resource-sharing does not depend on the kind of venue the resource is shared in. Unfortunately, many conference publications are not included in the major citation index or database (e.g., available in Web of Science, Scopus, or the Microsoft Academic Graph). Therefore, we expect that the data about these other kinds of events is scarcer.

Our work examined the essential dimension of longevity in resource sharing. To the best of our knowledge, it is one of the first investigations of this issue. Our work shows that technological factors are perhaps more critical than previously thought. In a sense, we speculate that the level of influence and prestige of the authors and institutions are not as important as the technology where the resources are shared. Of course, both factors are correlated—high-resource institutions will help store resources with better technologies. Still, our results might inform what institutions need to do with respect to sharing technologies. For example, we could add technological requirements for sharing information. In this sense, the guidelines provided by institutions such as the National Institutes of Health National Institues of Health (2020); National Institutes of Health, 2003) make this more explicit. The kinds of analysis we present here can support how these decisions are made in other institutions worldwide. Reproducibility and replicability hinge on our ability to share the resources in the best possible manner (Open Science Collaboration, 2015). The factors presented by our study shine light on what could be the steps needed to achieve this ideal.

Conclusion

The goal of this article was to investigate the factors associated with resource-sharing longevity and develop an automated resource expiration prediction to inform maintenance for research publications. We used a large sample of URLs from open access publications and associated them with the web page where they are shared, and the factors associated with publications that shared them. We found that the technology used by the sharer is an essential factor. Though significantly less important, the second set of factors was associated with the number of institutions citing the work sharing the resource instead of the author and journal numbers or prestige. We discussed some limitations, such as only being focused on biomedical publications and journals. Still, we found that our findings could be important for encouraging better technologies for storing resources.

We have many future avenues for exploration. The first is to better understand whether resources available are usable. For example, some resources might be available but contain incorrect data or code that does not run. This difference will require a significantly more sophisticated analysis of the *content* of the resources being shared. Also, we will expand the analysis to conferences, which we expect to suffer from more issues and have shorter longevity.

We hope that our results illuminate a critical component of reproducibility and replicability. Until now, most work has been encouraging resource sharing without regard to how long they are available. Our work helps the scientific community understand the factors that make longevity necessary. Also, it is striking how important technology is for longevity, which might be related to problems of equity around the world regarding access to the latest methods of sharing data. We hope that these issues get increasingly addressed as we achieve the mission of making resources outside paper first-class citizens in the production of knowledge.

Data availability

The data and code to reproduce the results are available at https://github.com/sciosci/predicting_resource_longevity.

Received: 11 February 2022; Accepted: 10 March 2025;
Published online: 22 May 2025

Notes

1 https://archive.org/web/

2 Our model also has a regularization parameter, and we can see how the parameter changes as the amount of regularization increases, a plot known as "regularization path" (Fig. S1, Supplementary Material)

References

Baker KS, Millerand F (2010) Infrastructuring ecology: challenges in achieving data sharing. In Collaboration in the New Life Sciences. Routledge

Bonàs-Guarch S, Guindo-Martínez M, Miguel-Escalada I, Grarup N, Sebastian D, Rodriguez-Fos E, Sánchez F, Planas-Fèlix M, Cortes-Sánchez P, González S (2018) Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. Nat Commun 9(1):1–14

Chen C, Luo B, Chiu K, Zhao R, Wang P (2014) The preferences of authors of Chinese library and information science journal articles in citing Internet sources. Libr Inf Sci Res 36(3):163–170

Duda JJ, Camp RJ (2008) 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

El-Bakry HM, Mohammed AA (2009) Optimal document archiving and fast information retrieval. Int J Comput Sci Eng 2:108–121

Fu H-Z, Chu J, Zhang M (2018) In-depth analysis of international collaboration and inter-institutional collaboration in nuclear science and technology during 2006–2015. J Nucl Sci Technol 55(1):29–40

Gazni A, Didegah F (2011) Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. Scientometrics 87(2):251–265

Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS (2020) Transparency and reproducibility in artificial intelligence. Nature 586(7829):E14–E16

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning (Vol. 112). Springer

Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P (2009) Data sharing in genomics—re-shaping scientific practice. Nat Rev Genet 10(5):331–335

Klein M, Nelson ML (2010). Evaluating methods to rediscover missing web pages from the web infrastructure. Proceedings of the 10th Annual Joint Conference on Digital Libraries, 59–68

Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, Tobin R (2014) Scholarly context not found: one in five articles suffers from reference rot. PloS One 9(12):e115253

Koehler W (1999) Digital libraries and World Wide Web sites and page persistence. Inf Res 4(4):4–4

Koehler W (2004) A longitudinal study of Web pages continued: A consideration of document persistence [Text]. Professor T.D. Wilson. http://informationr.net/ir/9-2/paper174.html

Kop M (2020) Machine Learning & EU Data Sharing Practices

Kulkarni AV, Aziz B, Shams I, Busse JW (2011) Author self-citation in the general medicine literature. PLoS ONE 6(6):e20885

Lawrence S, Giles CL (2000) Accessibility of information on the Web, intelligence, v. 11 n. 1. Spring

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67

Mangul S, Mosqueiro T, Duong D, Mitchell K, Sarwal V, Hill B, Brito J, Littman RJ, Statz B, Lam AK-M (2018) A comprehensive analysis of the usability and archival stability of omics computational tools and resources. *BioRxiv* https://doi.org/10.1101/452532

Mannheimer S, Pienta A, Kirilova D, Elman C, Wutich A (2019) Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. Am Behav Sci 63(5):643–664

Mbachu HI, Nduka EC, Nja ME (2012) Designing a pseudo R-squared goodness-of-fit measure in generalized linear models. J Math Res 4(2):148

Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, Koo B, Krishnakumar A, Biswal BB, Castellanos FX (2018) Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 9(1):1–7

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press

National Academies of Sciences & Medicine (2019) Reproducibility and replicability in science. National Academies Press

National Institues of Health (2020) Final NIH Policy for Data Management and Sharing. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

National Institutes of Health (2003) NIH Guide: Final NIH Statement On Sharing Research Date. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Foundation (2011) Dissemination and Sharing of Research Results|NSF—National Science Foundation. https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716

Penders B (2018) Ten simple rules for responsible referencing. Public Library of Science San Francisco, CA USA

Peng RD (2011) Reproducible research in computational science. Science 334(6060):1226–1227

Plantin J-C, Lagoze C, Edwards PN (2018) Re-integrating scholarly infrastructure: the ambiguous role of data sharing platforms. Big Data Soc 5(1):2053951718756683

Plavén-Sigray P, Matheson GJ, Schiffler BC, Thompson WH (2017) The readability of scientific texts is decreasing over time. Elife 6:e27725

Provos N, Mavrommatis P, Rajab M & Monrose F (2008) All your iframes point to us

Sampath Kumar BT, Prithvi Raj KR (2012) Availability and persistence of web citations in Indian LIS literature. Electron Libr 30(1):19–32. https://doi.org/10.1108/02640471211204042

Valcher ME (2015) The Value of Conferences [President's Message]. IEEE Control Syst Mag 35(4):10–11

Van de Sompel H, Klein M & Shankar H (2014) Towards robust hyperlinks for web-based scholarly communication. In S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, & J. Urban (Eds.), Intelligent Computer Mathematics (Vol 8543, pp 12–25). Springer International Publishing. https://doi.org/10.1007/978-3-319-08434-3_2

Van Horn JD, Gazzaniga MS (2013) Why share data? Lessons learned from the fMRIDC. Neuroimage 82:677–682

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM, Haendel MA (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ 1:e148

Wang K, Shen Z, Huang C, Wu CH, Eide D, Dong Y, Rogahn R (2019) A review of microsoft academic services for science of science studies. Front Big Data 2:45

Zeng T, Shema A & Acuna DE (2019) Dead science: Most resources linked in biomedical articles disappear in eight years. International Conference on Information, 170–176

Zeng T & Acuna DE (2020) Finding datasets in publications: the Syracuse University approach. In Rich Search and Discovery for Research Datasets (pp. 158–165). SAGE

Zhuang H, Huang TY, Acuna DE (2023) A computational analysis of accessibility, readability, and explainability of figures in open access publications. EPJ Data Sci 12(1):5

Zhou K, Grover C, Klein M & Tobin R (2015) No more 404s: Predicting referenced link rot in scholarly articles for pro-active archiving. Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, 233–236

Ziemann M, Eren Y, El-Osta A (2016) Gene name errors are widespread in the scientific literature. Genome Biol 17(1):177

Author contributions

DEA: Conceptualization, Methodology, Software, Writing—Original draft preparation, Supervision JJ: Conceptualization, Methodology, Software, Writing—Original draft preparation TZ: Conceptualization, Methodology, Software LL Writing—Original draft preparation HZ Writing—Original draft preparation.

Competing interests
The first author is a member of the editorial board of the journal. The remaining authors declare no compating interests.

Ethical approval
This article does not contain any studies with human participants performed by any of the authors.

Informed consent
This article does not contain any studies with human participants performed by any of the authors.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04716-z.

Correspondence and requests for materials should be addressed to Daniel E. Acuna.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.