JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 12, 1 (2024)

Digital Scholarly Journals Are Poorly Preserved: A Study of 7 Million Articles

Martin Paul Eve

Eve, M.P. (2024). Digital Scholarly Journals Are Poorly Preserved: A Study of 7 Million Articles. *Journal of Librarianship and Scholarly Communication*, 12(1), eP16288. https://doi.org/10.31274/jlsc.16288

This article underwent semi-anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.31274/jlsc.16288

RESEARCH ARTICLE

Digital Scholarly Journals Are Poorly Preserved: A Study of 7 Million Articles

Martin Paul Eve
Crossref and Birkbeck, University of London

ABSTRACT

Introduction: Digital preservation underpins the persistence of scholarly links and citations through the digital object identifier (DOI) system. We do not currently know, at scale, the extent to which articles assigned a DOI are adequately preserved.
Methods: We construct a database of preservation information from original archival sources and then examine the preservation statuses of 7,438,037 DOIs in a random sample.
Results: Of the 7,438,037 works examined, there were 5.9 million copies spread over the archives used in this work. Furthermore, a total of 4,342,368 of the works that we studied (58.38%) were present in at least one archive. However, this left 2,056,492 works in our sample (27.64%) that are seemingly unpreserved. The remaining 13.98% of works in the sample were excluded either for being too recent (published in the current year), not being journal articles, or having insufficient date metadata for us to identify the source.
Discussion: Our study is limited by design in several ways. Among these are the facts that it uses only a subset of archives, it only tracks articles with DOIs, and it does not account for institutional repository coverage. Nonetheless, as an initial attempt to gauge the landscape, our results will still be of interest to libraries, publishers, and researchers.
Conclusion: This work reveals an alarming preservation deficit. Only 0.96% of Crossref members (n = 204) can be confirmed to digitally preserve over 75% of their content in three or more of the archives that we studied. (Note that when, in this article, we write "preserved," we mean "that we were able to confirm as preserved," as per the specified limitations of this study.) A slightly larger proportion, i.e., 8.5% (n = 1,797), preserved over 50% of their content in two or more archives. However, many members, i.e., 57.7% (n = 12,257), only met the threshold of having 25% of their material in a single archive. Most worryingly, 32.9% (n = 6,982) of Crossref members seem not to have any adequate digital preservation in place, which is against the recommendations of the Digital Preservation Coalition.

Key words: digital preservation, persistent identifiers, scholarly communications

Received: 04/24/2023 Accepted: 09/20/2023

Corresponding author. Email: martin.eve@bbk.ac.uk (Martin Paul Eve)

IMPLICATIONS FOR PRACTICE

We suggest a number of actions that could strengthen preservation cultures, including the following:

1. DOI registration agencies upgrading their contractual wording from "best efforts" to define a minimum required standard of preservation, with certified archives listed. This will require periodic review.
2. DOI registration agencies beginning enforcement of the preservation clause, with appropriate member sanctions.
3. DOI registration agencies upgrading DOI deposit schemas to make preservation assertions mandatory.
4. DOI registration agencies and other groups, perhaps from the library community, regularly confirming the accuracy of preservation assertions.
5. Library groups and other organizations (such as the Open Access Scholarly Publishing Association [OASPA], the Directory of Open Access Journals [DOAJ], and the Library Publishing Coalition) creating and continuing an education and outreach campaign for publisher members that emphasizes digital preservation. This could also include webinars for new members.
6. DOI registration agencies conducting direct outreach to members in the lower-ranking preservation categories, bearing in mind that targeting larger members here may have disproportionate benefits.
7. DOI registration agencies considering nationally specific campaigns in which clear problems have emerged on a geographical basis.
8. DOI registration agencies introducing an opt-in, but automated, system for preserving content through DOI registration mechanisms.

We also note that the unavailability of item-level preservation data represents a significant challenge to observation of the digital preservation landscape.

INTRODUCTION

A crucial feature of the so-called scholarly record, which is difficult to define, but central also to the mission of academic libraries, is the incremental building of truths atop one another through citation (Dougherty, 2018). As Anthony Grafton sets out in his history of the footnote, "The culturally contingent and eminently fallible footnote offers the only guarantee we have that statements about the past derive from identifiable sources. And that is the only

ground we have to trust them" (Grafton, 1999, p. 233). Epistemic trust derives from being able to check the use of scholarly/scientific sources and to verify their claims. As the age-old adage goes, we see farther by standing on the shoulders of giants, or even of ordinary-sized people.

A vital aspect of this ecosystem for such validation is the persistence of addressing and the discoverability of scholarship in the digital era (Kenney et al., 2006, p. 5). Scholarly objects must be identified by a stable reference system if readers are to be able to find and check subsequent assertions based on them. For those working in universities, the academic library is the usual port of call for such discoverability. This originated in the era of purely print collections, in which the library was the custodian of all material. In our electronic age, when such systems include item locators (such as uniform resource locators [URLs]), the digital object identifier (DOI) system has become the *de facto* standard for providing stable and persistent scholarly addresses. Digital resources are also far more distributed and spread over servers worldwide rather than being localized in the on-campus library.

The DOI system, as implemented by Crossref, uses the HANDLE mechanism to dereference a DOI to a URL via a link resolver (Hendricks & Crossref, 2023). The fundamental concept of persistence in such a scheme lies in the ability to change the URL endpoint to which a DOI resolves. Therefore, for example, if a publisher goes bankrupt or if their URLs change, the address to which the DOI resolves can be updated to reflect a new location. In the event that the publisher ceases to exist, dark archives (which store copies of scholarly material behind the scenes and make it available on such a failure) such as CLOCKSS (Controlled Lots of Copies Keeps Stuff Safe), LOCKSS (Lots of Copies Keeps Stuff Safe), and Portico (among others) can open up the material and the DOI can be repointed. (These archives are called "dark" because their contents are only made accessible when the original source fails.) It is for this reason that Crossref was originally known as PILA (the Publishers International Linking Association). Ever-more scholarly articles now use DOIs as their primary persistent identifier (Gorraiz et al., 2016). Without a preservation system behind such a mechanism, there is a "gap […] between DOIs' potential and actual benefits" (Van de Sompel et al., 2016). However, this reliance on preservation reveals a fundamental aspect of DOIs (and other persistent identifiers): DOIs are enmeshed in social systems of publishing practice and are not merely a technical phenomenon.

Digital preservation consists of several different activities. As defined by the Digital Preservation Coalition, digital preservation "refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary." It is usually decomposed into time-limited periods:

> Short-term preservation: Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.
>
> Medium-term preservation: Continued access to digital materials beyond changes in technology for a defined period of time, but not indefinitely.
>
> Long-term preservation: Continued access to digital materials, or at least to the information contained in them, indefinitely (Digital Preservation Coalition, 2015).

In the case of the scholarly record, we desire long-term preservation so that knowledge claims may be sustained indefinitely into the future. Preservation activities cover everything from running server infrastructures to storing extra copies to ensuring that material is sent to archives at the time of publication.

However, if the stability and persistence of the DOI linking mechanism is to be trusted, a fundamental question must be addressed: What proportion of scholarly objects assigned a DOI are adequately preserved in a recognized dark archive? Preservation remains a crucial support for persistent identification (see Bogdanski, 2006). It has, in fact, been described as the "grand challenge" of our time for academia (Case, 2016). How stable are the preservation systems that underpin the persistence of persistent identifiers? And which actors are behaving well on this theme? Given the scale of the problem, we might even ask, as did Peter Burnhill and Lisa Otty in 2015, whether it is already "too late to ensure continuity of access to the scholarly record?" (Burnhill and Otty, 2015).

Remarkably, we do not know the answer to these questions, despite previous attempts at quantitative analysis (Seadle, 2011; Burnhill & Otty, 2015, p. 17). Certainly, surveys of the academic library community have noted that this is a problem of the utmost significance, even while most libraries are not actively engaged in preservation efforts (Meddings, 2011, p. 57; see also Salo, 2020). The Crossref schema recognizes this importance and allows metadata depositors to assert that content is preserved and to specify the archive in which material appears (Hendricks, 2023). However, the truth of these preservation assertions has not previously been tested. In this article, we set out a method for answering this question and present our findings across a representative sample of Crossref DOIs.

That said, an important point to make is that there is no consensus over who should be responsible for archiving scholarship in the digital age (Anderson, 2012). Some studies, on the one hand, have assumed that this will continue to be a function of the academic

library, as it was when libraries had physical custody of the objects in question (Dressler, 2017). Indeed, the LOCKSS system operates its nodes out of academic libraries. It is also true that many libraries are also publishers (Moulaison & Million, 2015). On the other hand, it has become the responsibility of publishers to ensure that "their" content is preserved, doubtless a legacy of the copyright transfer systems that have prevailed under the subscription-access model for scholarly communications. At least part of the reason for this is that the contractual terms of DOI assignment, at least at Crossref, is an agreement on the part of the member to "use best efforts to contract with a third-party archive or other content host (an "Archive") […] for such Archive to preserve the Member's Content and, in the event that the Member ceases to host the Member's Content, to make such Content available for persistent linking" (Hendricks & Crossref, 2022). Despite the controversy surrounding who archives, in this article we assume that such responsibility rests with publishers of academic material, while acknowledging that this could be done differently.

The threat posed by the disappearance of the scholarly record is real (Jamali et al., 2022). Without active understanding and intervention, we will continue to lose valuable material and threaten the persistence of digital links to scholarship and research.

METHODS

The closest database to a centralized canonical source of preservation data for scholarly journals is The Keepers registry (Burnhill, 2009; Burnhill & Guy, 2010). Originally a pilot undertaken by the United Kingdom's Joint Infrastructure Services Council (Jisc) and Edina, the database has since been moved to the International Standard Serial Number (ISSN) International Centre ("Enhancing The Keepers Registry," 2016). According to the ISSN International Centre, The Keepers registry has three core purposes:

- To enable librarians and policy makers to find out who is looking after what e-content, how, and with what terms of access.
- To highlight e-journals that are still "at risk of loss" and need to be archived.
- To showcase the archiving organizations around the world, i.e., The Keepers, which provide the digital shelves for access to content over the long term (ISSN International Centre, 2023a).

However, The Keepers registry has several fundamental flaws that make it unsuitable for a research project of this scale:

1. There is no programmatic application programming interface (API) access. Systematically checking many hundreds of thousands of records using a manual Web interface would be prohibitively slow.
2. The licensing page from the ISSN International Centre explicitly prohibits such research uses, noting that "any use of ISSN Data, other than those expressly authorised above, is prohibited, including the following uses: reproducing, disseminating to unauthorised third parties or making publicly accessible ISSN Data which is reusable and/or established in a format readable by anyone, including on the Internet" (ISSN International Centre, 2023b).

Therefore, it was necessary for us to rebuild the preservation database from scratch/the original sources (see also Galyani Moghaddam, 2008; Mering, 2015).

Although there is no consensus as to what constitutes a valid archive, data sources used by The Keepers registry are as follows:

- Archaeology Data Service
- British Library
- Cariniana Network
- CLOCKSS Archive
- Gallica
- HathiTrust
- Internet Archive
- Library of Congress
- LOCKSS Archive
- Merritt Preservation Repository
- National Digital Preservation Program, China
- National Library of the Netherlands
- Portico
- Public Knowledge Project PLN
- Scholars Portal
- Swiss National Library
- ZBW - Leibniz Information Centre for Economics

Some of these archives are very small. For instance, the Archaeology Data Service preserved just 10 items at its last update in 2020. Similarly, the ZBW - Leibniz Information Centre for Economics had just 35 preservation records. The Merritt Preservation Repository and the Swiss National Library also held just a few hundred records each. By contrast, the largest archive at the time of writing was Portico, which housed 32,835 records.

The Keepers registry also provides some statistics about their collection that can help in the prioritization of different archives. Specifically, they report the number of unique records per archive (i.e., titles preserved in just a single archive), as shown in Table 1. In other words, as an example, there are 10,621 titles, according to The Keepers registry, preserved only in HathiTrust.

Archive	Unique Titles	Catalogue Data Publicly Available?
Gallica	16,800	No
HathiTrust	10,621	Yes
Portico	8,231	Yes
Internet Archive	6,532	Yes
Public Knowledge Project PLN	2,859	Yes
Scholars Portal	1,773	No
CLOCKSS	1,762	Yes
Cariniana	1,543	Yes

Table 1. Data sources with unique titles and data availability.

All other archives had fewer than 1,000 unique records each. For each of these larger archives, we wrote an importer that would ingest these catalogues directly from the original source and allow us to resolve whether an item is preserved (Eve & Crossref, 2023b). Notably, most of the preservation data that we ingest describes content at the container level rather than the item level. Each archive also presents this description in an esoteric fashion, with little standardization among systems (Burnhill, 2013, p. 9–10). Therefore, to understand whether an item with a DOI is preserved requires a resolution of the metadata/dereferencing a DOI and then comparing these data with the container-level information.

One of the archives here listed posed challenges. Gallica, although listed as a source for The Keepers registry, does not appear to publish downloadable dumps of serials data. Despite contacting the French National Library, we received no response and thereby omitted Gallica from our database. Table 2 shows the eventual source list that we were able to use.

Archive	File	Test Date
Cariniana	http://reports-lockss.ibict.br/keepers/pln/ibictpln/keepers-IBICTPLN-report.csv	2023-02-17
CLOCKSS	https://reports.clockss.org/keepers/keepers-CLOCKSS-report.csv	2023-02-16
HathiTrust	https://www.hathitrust.org/hathifiles	2023-02-23
Internet Archive/ FATCAT	https://archive.org/details/ia-keepers-registry-kbart	2023-02
LOCKSS	https://reports.lockss.org/keepers/keepers-LOCKSS-report.csv	2023-02-20
PKP PLN	https://pkp.sfu.ca/files/pkppn/onix.csv	2023-02-20
Portico	https://api.portico.org/kbart/Portico_Holding_KBart.txt	2023-02-21
Scholars Portal	Private data source. Please contact archive.	2023-02-21

Table 2. Sources used.

Future work could consider other networks and platforms. For instance, Figshare is backed by the Chronopolis digital preservation system at the University of California at San Diego. For this study, looking only at journal articles, Figshare is perhaps not a major archive, specializing as it does in in data deposit. However, this is a good example of an open repository that nonetheless holds a preservation function. We acknowledge that this study was only able to cover a smaller subset of preservation archives than a real-world environment would use.

Data validation, to ensure that the data we ingested matched that in The Keepers registry, and that we translated correctly from item-level to container-level metadata, was performed against 20 DOIs from each archive, selected at random using our tool's random-sample function. For each archive, we checked five containers that should yield a positive result and five containers that should yield a negative. For each container, we checked a minimum of two DOIs, selected at random by hand. We then manually compared each title's records with The Keepers registry to ensure that our records matched. A full record of these checks is available in the database's code repository (Eve & Crossref, 2023c).

In order to appraise practice in this space, we also needed a set of benchmarks by which to grade members. There is no prevailing code of practice at present for the number of preservation instances/archives that should be used. However, it makes sense to assume that content should be preserved in more than one archive. If not, then, if that archive fails (for institutional-social or technical reasons), there is no secondary redundancy. We settled, in the end, on a set of criteria that reflect the proportion of a Crossref member's output that is preserved in a specific number of archives. This resulted in the taxonomy shown in Table 3.

Grade	Criteria
Gold	Gold members are those that have 75% of their content digitally preserved in three or more recognized archives.
Silver	Silver members are those that have 50% of their content digitally preserved in two or more recognized archives.
Bronze	Bronze members are those that have 25% of their content digitally preserved in one or more recognized archives.
Unclassified	Unclassified members are those that do not meet the above criteria.

Table 3. The grading system used to group members.

Finally, once we had built this database, the next step was to compare this against Crossref members. However, as of April 2023, the Crossref metadata database contains over 144 million records (Crossref, 2023). To sequentially dereference all of these records and then to compare preservation information would have taken a prohibitively long time. Therefore, instead, we extracted 1,000 random sample DOIs from each member using the Crossref sampling framework (or fewer if the total number of records was less than 1,000) (Tkaczyk et al., 2023). Samples are stored in the public Amazon AWS S3 bucket "samples.crossref.org." The data that underpin this paper were collected using the samples in the "members-works/2023-03-05/samples/" keys. This led to a total appraisal of 7,438,037 DOIs.

RESULTS: THE PRESERVATION LANDSCAPE

The initial overview of Crossref members, shown in Figure 1, reveals a scholarly landscape with an imperilled digital future (although things have improved over the last decade; see Rieger & Wolven, 2011). Only 0.96% of Crossref members (n = 204) could be detected to preserve over 75% of their content in three or more of the archives that we studied. A slightly larger proportion, 8.5% (n = 1,797), seemed to preserve over 50% of their content in two or more archives. However, many members, 57.7% (n = 12,257), only met the threshold of having 25% of their material in a single archive that we could detect. Most worryingly, 32.9% (n = 6,982) of Crossref members seem, using our data set, not to have any adequate digital preservation in place, which is against the recommendations of the Digital Preservation Coalition (Beagrie, 2013, p. 33).

Of course, examining preservation at the member level only tells part of the story. Looking at the breakdown across actual works is also helpful. In the 7,438,037 works examined, there were 5,913,102 "preservation instances." This is a term denoting the number of stored copies. Therefore, a single work that is preserved in three archives has three "preservation instances." For example, if I examined three works total, and one of them was stored in three archives

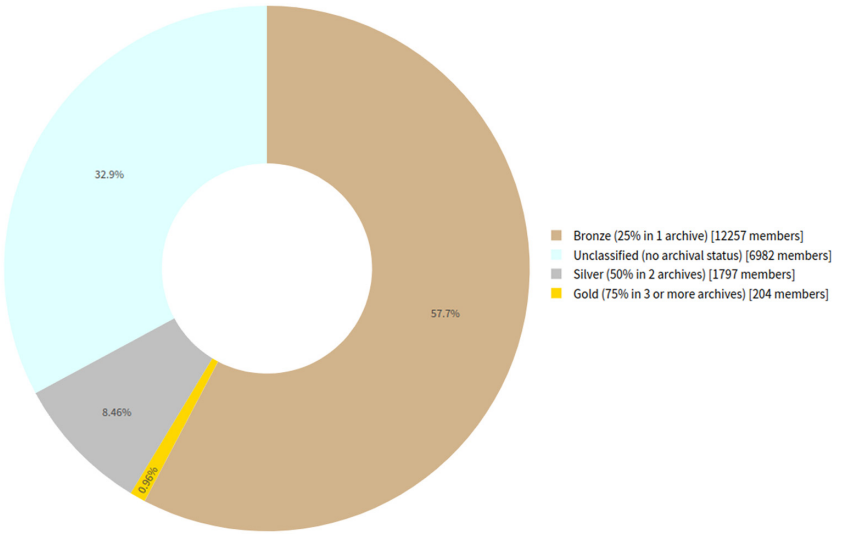

Figure 1. The percentages of Crossref members in each preservation category (for higher resolution/ interactive charts of all figures, please see Eve & Crossref, 2023a).

whereas the other two were stored in no archives, there would be a total of three preservation instances. Furthermore, a total of 4,342,368 of the works that we studied (58.38%) did have at least one preservation instance. However, this left 2,056,492 works in our sample (27.64%) that seem unpreserved. The remaining 13.98% of works in the sample were excluded either for being too recent (published in the current year), not being journal articles, or having insufficient date metadata for us to identify the source.

Another question that we can address from this data set is as follows: Which categories of Crossref members do things well? And which have room for improvement?

Crossref members can be categorized according to a variety of features that can help us to understand trends among different publisher types. The "size" of a member, for example, can be determined by their fee band, which correlates to revenue from publishing. However, it can also be specified by number of deposits that a member has made. We also have data on the country of origin for a publisher, which, although needing caution, can reveal geographical cultural norms around preservation practices and knowledge. The next sections break down our findings into these category buckets.

Preservation by member revenue

When we plot "gold" members (those with 75% or more stored in three or more archives), as shown in Figure 2, there is little variance among the sizes of publisher who achieve this score,

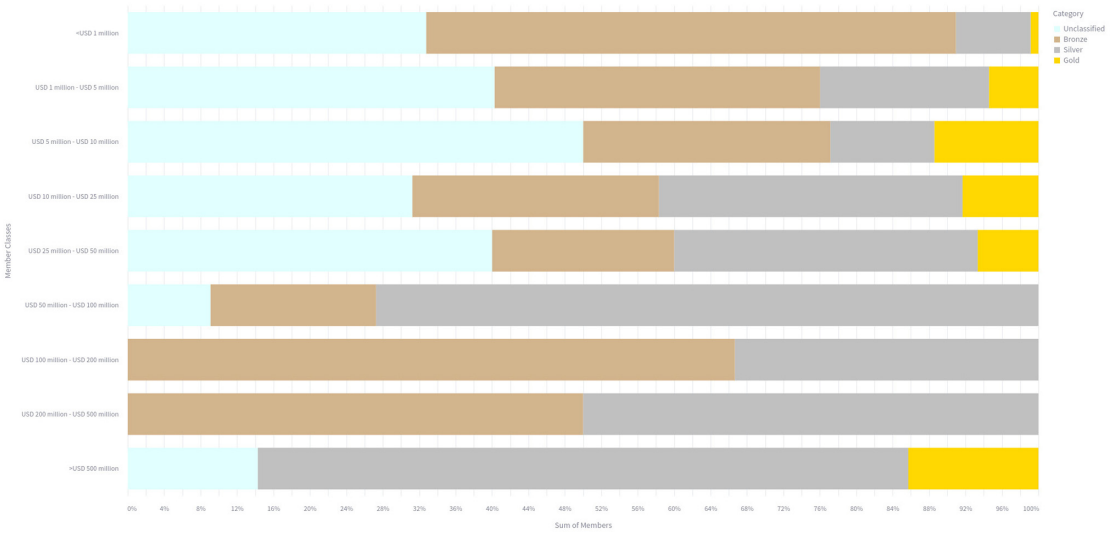

Figure 2. Proportion of members in each fee band that score in each of the preservation categories.

except for the very small publishers. Although we might expect well-resourced publishers in the highest revenue category to have the best digital preservation practices, only one of the largest members (Elsevier) scored in this category. Meanwhile, "smaller" members (even those with publishing revenues of $50 million USD) fare worse. Finally, publishers with less than $1 million USD of publishing revenue rarely have the highest level of robust digital preservation. That said, because we are dealing with proportions of members, and because there are many more smaller members than larger publishers, it takes only a small variance in the number of larger publishers to change their ratio substantially (see Table 4).

Membership Category	Number of Members
<1 million USD	20,863
1 million to 5 million USD	221
5 million to 10 million USD	70
10 million to 25 million USD	48
25 million to 50 million USD	15
50 million to 100 million USD	11
100 million to 200 million USD	3
200 million to 500 million USD	2
>500 million USD	7

Table 4. Crossref membership categories and the number of members in each.

Publishers in the "silver" category, i.e., with 50% of content preserved in two or more archives, make up a much larger proportion of all publishers. Of the seven mega-publishers in the highest fee bracket at Crossref, five meet the criteria for "silver." Again, the lowest proportion of members in this category come from those publishers with under 1 million USD of publishing revenue.

When it comes to "bronze" members, i.e., a very minimal standard for digital preservation, in which 25% of published works are stored in at least one archive, the picture is inverted. Apart from the single (n = 1) skewing outlier in the "200 million to500 million USD" category, it is clear that many smaller and mid-size publishers are operating at this low level of preservation. Indeed, in the <1 million USD category, a total of 12,138 members had content preserved only at this potentially less safe level.

The final category, i.e., members who have inadequate digital preservation procedures, is "unclassified." This consists of members who do not meet any of the aforementioned criteria. Of course, it is possible that such members have taken preservation precautions that are not covered by our database and analysis (see the Limitations section). The proportion of members in this category increases substantially as we reach the lower/smaller fee banding levels.

The most curious outlier here is the single member in the highest fee category that appears to have no digital preservation: Clarivate. Although it is possible that Clarivate has preserved print correlate publications by national legal deposit mechanisms, The Keepers registry also holds no data on preservation for the sample that we checked.

However, as might be expected given the previous analyses, many smaller members, when categorized by fee size, do not fare so well at digital preservation. A large number of the smallest category of members (n = 6,835) seem to have no adequate digital preservation.

A clear, and perhaps obvious, trend emerges from this analysis of publisher revenues correlated against preservation: Wealthier publishers are likely to have (but do not unambiguously have, in all cases) better digital preservation than less well-resourced Crossref members. Many smaller members seem not to have adequate digital preservation, and their content is at risk.

Preservation by member deposit numbers

However, wealth and revenue are not the only way to judge a Crossref member by size. It is possible to have low levels of revenue but to nonetheless deposit a large number of DOIs. Therefore, it is also worth examining member deposit levels when accounting for number of deposits.

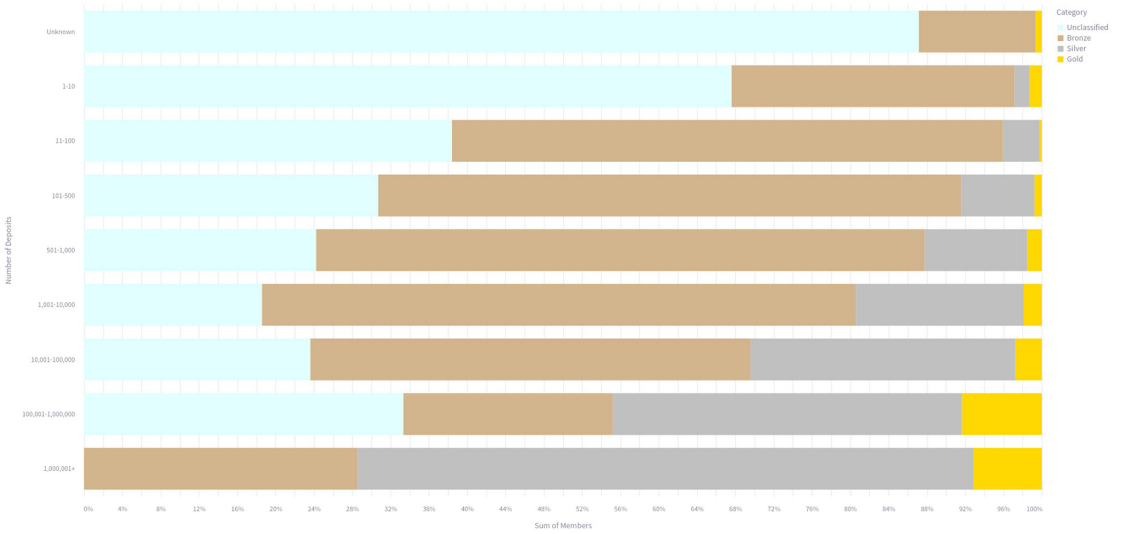

Figure 3. Proportion of members in each deposit-level band that score in each of the preservation categories.

Figure 3 shows that similar proportions of memberships are preserving at the gold level, across all sizes of deposit category. Again, however, of the largest publishers, only one meets this high threshold of preservability. Meanwhile, a set of mid-size members fare better, before we see a tail-off again at the smaller end of the scale.

As with the aforementioned revenue categories, at the "silver" standard, there is a clear skew toward mid-to-large size publishers meeting this benchmark. As with the revenue figures, the largest publishers nearly all meet this level of preservation, but there are not many of them. Meanwhile, a large number of smaller publishers are not meeting this standard.

As we move down the grading system, into the "bronze" category, it becomes clear, once more, that many more publishers toward the smaller end fall into this bracket. Finally, the number of members in the "unclassified" category shows a similar distribution to bronze members, growing substantially in proportionate size for the smallest members.

This grouping by number of deposits tells a similar story to the aforementioned revenue-based chart: The majority of larger publishers achieve a higher standard of digital preservation than many of their smaller counterparts. However, this is not the full picture. Clearly, a total of 1,000,000 deposits at 75% is a substantially more significant preservation portion than the same 75% of a member who only has 100 deposits. Therefore, any remedial action or intervention to encourage preservation practices should consider each member's output volume and the additional quantitative number of works that will be preserved as a result of that intervention.

National preservation cultures

Another taxonomy by which we can group Crossref members is their country of origin. Although care must be taken in such instances to avoid national stereotyping, it is useful to know whether there are specific geographical regions with better or worse preservation practices. This knowledge will allow for targeted intervention, in local languages, that are attuned to the specific cultural-geographic features of these regions. It is also true that preservation cultures vary among regions due to varying legal deposit laws (amid other factors) (see, e.g., Choi & Park, 2007).

As in the preceding sections, Figure 4 shows members, this time grouped by country. Figure 4 is merely an excerpt of the full figure, which is available in the data set because the whole chart is too large. The graph is normalized to show a percentage of member from each region. Once more, this leaves the visualization open to misinterpretation. Tortola, for instance, only has two members. Therefore, to achieve its seeming outlier status of 50% of members being gold only requires a single member to hold this status. Meanwhile, the United States, although dwarfed by Tortola's portion in the visualization, actually has 91 members in this category.

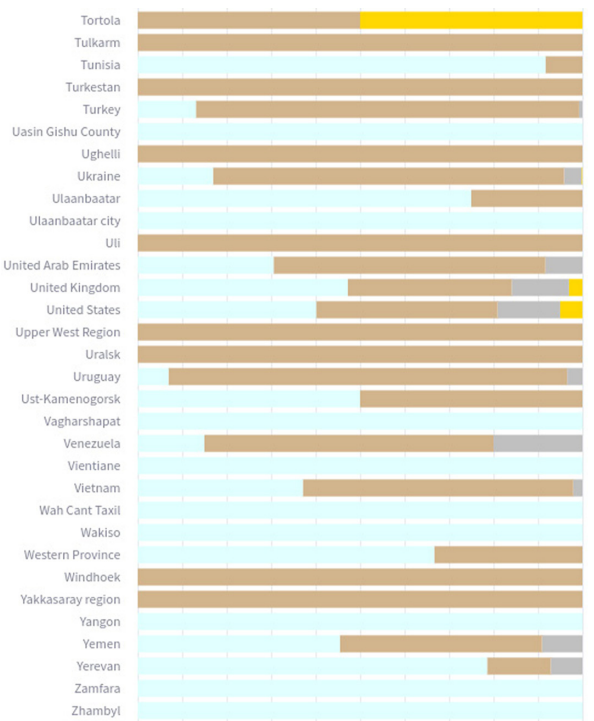

Figure 4. A sample of the breakdown by country figures. As with all figures, the full chart and visualizations are available at the following link: https://the-vault.fly.dev/.

Similarly, for "silver" members, several smaller nations achieve 100% of members at this level. Again, this is because all of the nations in question (St Michael, etc) have only one member. This silver set also reveals some flaws and inconsistencies in the metadata used for the study. In this data set, some members are missing the country name when extracted. As a result, cities such as "Ghauri Town Phase 5" are listed instead of "Pakistan." This is a result of nonstandardized metadata for location information in Crossref's database, which should be fixed in future updates on this work.

When we reach the two lower grades, i.e., bronze and unclassified, these problems of data extraction are exacerbated, with many more misidentified countries. For instance, there is no clear pattern to be discerned among these data, and the entire picture is muddied by low-count nations where small variances in absolute figures cause large fluctuations in the proportion, up to 100%. The same can be said for unclassified nations.

In short, although nationhood may be a useful dimension to consider when making targeted preservation interventions, it is hard to discern any distinct pattern that identifies national characteristics that might lead to good preservation outcomes. There is a straightforward correlation with neither level of economic development nor size of country. Therefore, any interventions that wish to take advantage of national characteristics must simply work on a case-by-case basis, with which the interactive data set that accompanies this article may help. For instance, a relatively low portion of the 2,879 members from Indonesia are achieving the highest of preservation standards (as, say, they are also in the United States). This could merit a specific national campaign, in local languages and with local cultural understanding, to boost preservation in such countries.

Interactive data set, data availability, and software

The data set that underpins this article can be accessed in two distinct ways. First, a visualization and data presentation site ("The Vault") is available for consultation (Eve & Crossref, 2023a). This site includes an overall report, per-member data, member class data by size, a log of the data validation process that we used to check the sources, and a description of the grading system.

Alternatively, the raw data are available in the public AWS S3 bucket: outputs.research.crossref.org. Data are stored in the "annotations" folder/key, under both works and members. The "reports" folder/key contains the aggregated data set used in this article.

There are two software components that were used to generate the data for this project: the preservation database and the preservation reporting Streamlit site. The first of these is a

command-line python utility that allows for the building of the preservation status relational database and the ability then to look up a DOI against that database and to generate the overall report data (Eve & Crossref, 2023b). The second software component marshals the data into Pandas DataFrames and then uses Streamlit to display interactive visualizations of these data (Eve & Crossref, 2023d). Together, these components will allow the reader to reproduce the findings of this article.

DISCUSSION AND LIMITATIONS

There are a number of limitations to this study that should be fully appreciated before reaching any conclusions. The first and most important limitation is that this methodology under-counts actual preservation. We have only been able to include the catalogues of the most prominent dark archives, and we do not account for print correlate preservation. That is, it is possible that material that seems "unpreserved" by our counts does actually sometimes have a print version stored in, say, a national deposit library. This limitation is mitigated by the fact that it is better to be cautious than overly bold about the state of digital preservation.

It is also true that much content is "preserved," often illegally, in shadow libraries/archives such as Library Genesis and Sci-Hub (for more, see Bodó, 2018a, 2018b; Eve, 2022). These archives are at great legal threat of shutdown but have also proved surprisingly resilient. We do not, in this article, count material stored in such archives, even though it constitutes an additional storage location. Likewise, "green" open-access deposit (often of authors' accepted manuscripts) in institutional repositories has some preservation characteristics for which we do not account. However, comprehensively matching these deposits to the identifier at the version of record cannot accurately be performed at scale. It is also possible that some journals that we did not include are preserving material under their own steam, but the extent of such a practice remains nearly impossible to measure at scale.

The analysis in this paper is also only as good as the metadata that Crossref holds on a particular publication, and it only covers material that has been assigned a DOI by Crossref. Therefore, there is likely a long-tail of published material in smaller journals, not using DOIs, for which we are unable to determine preservation information (see also Rosenthal, 2020). As shown in the software source code, we employ a number of matching strategies to dereference a DOI to container-level data. However, some works are missing date information, whereas others are missing volume data, etc. Therefore, we have exposed the data set for every member so that we can accept corrections when errors are discovered. In some instances of incomplete or incorrect metadata, we are unable to match a DOI to preservation data, again leading to an under-count.

Future work may also wish to consider whether there are other fruitful groupings for trend analysis that we have not been able to cover at this time. A good example of this is, perhaps, disciplinary breakdown, which would require an accurate classification system for works and journals. Finally, non-standardized country and location information for publishers has hindered our ability to analyze members by region. There are also unresolved questions around what a publisher's "location" denotes: This can mean where they are meaningfully based or, alternatively, where they are incorporated (say, for tax reasons).

CONCLUSIONS AND RECOMMENDATIONS

In 2005, almost two decades ago, Don Waters, the Senior Program Officer for Scholarly Communications at the Andrew W. Mellon Foundation edited a consensus statement in the Association of Research Libraries newsletter, titled "Urgent Action Needed to Preserve Scholarly Electronic Journals" (Waters, 2005; see also Barnes, 1997; Cantara, 2004; and Waters, 2007). Many of the calls in that piece were heeded; we have archives that can provide the minimum level of service described therein and a comprehensive persistent identifier scheme on top of this (see, for instance, Day, 1998). Recent efforts such as Project JASPER have also highlighted the importance of preservation in the brave new world of open-access publishing ("Project JASPER," 2020).

However, as this article and its data show, the state of digital preservation of serials remains fragile in 2023, and these calls have not fully been answered (for another example of such a call, see Reich, 2006). This should be a substantial concern to the academic library community, whose mission remains to provide ongoing access to and discoverability of scholarship and research. A significant portion, approximately 28%, of academic journal articles with DOIs appear entirely unpreserved in the archives that we sampled, endangering both persistent identifier systems and the chain of verifiable citation that they are meant to underwrite. This confirms the findings of other studies that have examined the disappearance of open-access journals (Lightfoot, 2016; Laakso et al., 2021). It is also, of course, a problem not confined merely to academic journals; the digital preservation of all electronic resources poses challenges (Ainsworth et al., 2015). Availability of material, i.e., the aspect of preservation studied in this article, is also not the be-all and end-all. Other preservation concerns include the very real threat of format obsolescence, as just one example (Rosenthal, 2007, 2010; see also Wittenberg et al., 2018). Indeed, digital preservation is an ongoing activity, not a one-time deposit, that requires constant re-investment, reinvention, and labor (Cramer et al., 2023). In the coming years, the importance of considering, also, the environmental impacts of preservation strategies will be of importance (Pendergrass et al., 2019).

Although there are clearly some patterns, such as the fact that most of the ultra-large publishers have strong preservation cultures and practices, it is hard to explain this entirely. In recent years, the emergence of the Public Knowledge Project's Private LOCKSS Network has made preservation accessible even to those without the economic resources that are required to participate in, say, CLOCKSS and Portico (Sprout & Jordan, 2018). This circumvents the problem that most library respondents had in a recent survey on digital preservation: that the additional costs would be unaffordable (Meddings, 2011, p. 59). National preservation networks such as Cariniana can also help to address regionally specific challenges (Arellano, 2021). The fact that these networks are available, but not universally used, indicates that there is an awareness and education dimension to the preservation conundrum. Direct integration with archives in open-source journal management systems can also help to foster good practice (for example, Public Knowledge Project's Open Journal Systems integrating with their LOCKSS network or Janeway's plugin for Portico deposit) (Byers & Sanchez Lopez, 2023). However, archival deposit mechanisms are often antiquated, using frustrating older protocols such as file transfer protocol (FTP) or relying on the archive itself harvesting material on a period crawl.

It is also important to note that digital preservation overlaps the terrains of business/economics, social practices, and technical implementations (Reich & Rosenthal, 2009). It is not enough to propose technical solutions. Solid digital preservation implementations require social and business changes to publishers' practices. Certainly, technology can drive such changes. But without a core understanding of preservation principles and a commitment to persistence, change is unlikely to be meaningful. Preservation practices must be socially embedded at the heart of publisher activities.

Other studies have already suggested the range of ways in which preservation uptake could be improved (see Regan, 2016). However, there are a range of actions and activities that persistent identifier providers (and others) could undertake to promote better preservation practices:

1. DOI registration agencies upgrading their contractual wording from "best efforts" to define a minimum required standard of preservation, with certified archives listed. This will require periodic review.
2. DOI registration agencies beginning enforcement of the preservation clause, with appropriate member sanctions.
3. DOI registration agencies upgrading DOI deposit schemas to make preservation assertions mandatory.
4. DOI registration agencies and other groups, perhaps from the library community, regularly confirming the accuracy of preservation assertions.

5. Library groups and other organizations (such as OASPA, DOAJ, and the Library Publishing Coalition) creating and continuing an education and outreach campaign for publisher members that emphasizes digital preservation. This could also include webinars for new members.
6. DOI registration agencies conducting direct outreach to members in the lower ranking preservation categories, bearing in mind that targeting larger members here may have disproportionate benefits.
7. DOI registration agencies considering nationally specific campaigns in which clear problems have emerged on a geographical basis.
8. DOI registration agencies introducing an opt-in, but automated, system for preserving content through DOI registration mechanisms.

Although preservation deficits are not likely to be resolved in the very near future, taking action now will improve the situation and help to safeguard the digital scholarly record.

ACKNOWLEDGMENTS AND CONFLICT OF INTEREST STATEMENT

The author, Martin Paul Eve, works for Crossref. The author wishes to thank Jefferson Bailey, Geoffrey Bilder, Esha Datta, Rachael Lammey, Bryan Newbold, Ernesto Priego, Dorothea Salo, Dominika Tkaczyk, and Joe Wass for help with this work.

REFERENCES

Ainsworth, S. G., Nelson, M. L., & Van de Sompel, H. (2015). Only one out of five archived web pages existed as presented. *HT'15: Proceedings of the 26th ACM Conference on Hypertext & Social Media*, 257–266. https://doi.org/10.1145/2700171.2791044

Anderson, R. (2012 March 7). *E-journal preservation and archiving: Whether, how, who, which, where, and when?* The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2012/03/07/e-journal-preservation-and-archiving-whether-how-who-which-where-and-when/

Arellano, M. A. M. (2021 November 3). *The Cariniana network for digital preservation*. Digital Preservation Coalition. https://www.dpconline.org/blog/wdpd/cariniana-wdpd21

Barnes, J. (1997). Electronic archives: An essential element in complete electronic journals solutions. *Information Services & Use 17*(1), 37–47. https://doi.org/10.3233/ISU-1997-17105

Beagrie, N. (2013). *Preservation, trust and continuing access for e-journals*. Digital Preservation Coalition. https://doi.org/10.7207/twr13-04

Bodó, B. (2018a). Library genesis in numbers: Mapping the underground flow of knowledge. In J. Karaganis (Ed.), *Shadow libraries: Access to educational materials in global higher education*. pp. 53–78. The MIT Press.

Bodó, B. (2018b). The genesis of library genesis: The birth of a global scholarly shadow library. In J. Karaganis (Ed.), *Shadow libraries: Access to educational materials in global higher education*. pp. 25–52. The MIT Press.

Bogdanski, E. L. (2006). Serials preservation at a crossroads. *Serials Review*, *32*(2), 70–72. https://doi.org/10.1080/00987913.2006.10765033

Burnhill, P. (2009). Tracking e-journal preservation: Archiving registry service anyone? *Against the Grain*, *21*(1). https://doi.org/10.7771/2380-176X.2496

Burnhill, P. (2013). Tales from The Keepers registry: Serial issues about archiving & the web. *Serials Review*, *39*(1), 3–20. https://doi.org/10.1080/00987913.2013.10765481

Burnhill, P., & Guy, F. (2010). Piloting an e-journals Preservation Registry Service (PEPRS). *The Serials Librarian*, *58*(1–4), 117–26. https://doi.org/10.1080/03615261003622742

Burnhill, P., & Otty, L. (2015). *Is it too late to ensure continuity of access to the scholarly record?* [Conference Presentation] Proceedings of the IATUL Conferences, Hanover, Germany. https://docs.lib.purdue.edu/iatul/2015/ddp/6

Byers, A., & Sanchez Lopez, M. (2023). Portico [Python plugin]. Birkbeck Centre for Technology and Publishing. https://github.com/BirkbeckCTP/portico

Cantara, L., ed. (2004). *Archiving electronic journals*. Digital Library Federation. https://old.diglib.org/preserve/ejp.htm

Case, M. (2016). Preservation and scholarly communication: The grand challenges of our time. *Technicalities*, *36*(5), 3–6.

Choi, H. N., & Park, E. G. (2007). Preserving perpetual access to electronic journals: A Korean consortial approach. *Library Collections, Acquisitions, & Technical Services*, *31*(1), 1–11. https://doi.org/10.1080/14649055.2007.10766142

Cramer, T., German, C., Jefferies, N., & Wise, A. (2023). A perpetual motion machine: The preserved digital scholarly record. *Learned Publishing*, *36*(2), 312–318. https://doi.org/10.1002/leap.1494

Crossref. (2023 April 5). Crossref Stats. Crossref. https://www.crossref.org/06members/53status.html

Day, M. W. (1998). Online serials: Preservation issues. *The Serials Librarian*, *33*(3–4), 199–221. https://doi.org/10.1300/J123v33n03_01

Digital Preservation Coalition. (2015). *Digital Preservation Handbook: Glossary.* 2nd edition. Digital Preservation Coalition. https://www.dpconline.org/handbook/glossary

Dougherty, M. V. (2018). Defining the scholarly record. In M. V. Dougherty (Ed.), *Correcting the scholarly record for research integrity: In the aftermath of plagiarism* (pp. 19–57). Research Ethics Forum book series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99435-2_2

Dressler, V. A. (2017). The state of affairs with digital preservation at ARL member libraries: A survey and analysis of policy. *Digital Library Perspectives*, *33*(2), 137–55. https://doi.org/10.1108/DLP-08-2016-0030

Enhancing The Keepers Registry. (2016). Jisc. 2016.

Eve, M. P. (2022). Lessons from the library: Extreme minimalist scaling at pirate ebook platforms. *Digital Humanities Quarterly*, *016*(2).

Eve, M. P., & Crossref. (2023a). Preservation Report. The Vault. https://the-vault.fly.dev/

Eve, M. P., & Crossref. (2023b February 13). Crossref Labs Preservation-Database. GitLab. https://gitlab.com/crossref/labs/preservation-database

Eve, M. P., & Crossref. (2023c February 17). 2023-02 - Data Check. GitLab. https://gitlab.com/crossref/labs/preservation-data/-/blob/main/2023-02%20-%20Data%20Check.md

Eve, M. P., & Crossref. (2023d April 6). Crossref Labs Preservation-Reporting. GitLab. https://gitlab.com/crossref/labs/preservation-reporting

Galyani Moghaddam, G. (2008). Preserving scientific electronic journals: A study of archiving initiatives. *The Electronic Library*, *26*(1), 83–96. https://doi.org/10.1108/02640470810851761

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008

Grafton, A. (1999). *The footnote: A curious history*. Harvard University Press.

Hendricks, G. (2023). *Archive locations*. Crossref. https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/archive-locations/

Hendricks, G., & Crossref. (2022). Membership Terms. Crossref. https://www.crossref.org/membership/terms/

Hendricks, G., & Crossref. (2023). About Us. Crossref. https://www.crossref.org/community/about/

ISSN International Centre. (2023a). About Keepers registry. The Keepers registry. https://keepers.issn.org/keepers-registry

ISSN International Centre. (2023b). License Contract. *ISSN Portal*. https://portal.issn.org/content/license-contract

Jamali, H. R., Wakeling, S., & Abbasi, A. (2022). Why do journals discontinue? A study of Australian ceased journals. *Learned Publishing*, *35*(2), 219–228. https://doi.org/10.1002/leap.1448

Kenney, A. R., Entlich, R., Hirtle, P. B., McGovern, N. Y., & Buckley, E. L. (2006). E-journal archiving metes and bounds: A survey of the landscape. CLIR Publication, no. 138. Council on Library and Information Resources, Washington, DC. https://www.clir.org/pubs/reports/pub138/

Laakso, M., Matthias, L., & Jahn, N. (2021). Open is not forever: A study of vanished open access journals. *Journal of the Association for Information Science and Technology*, *72*(9), 1099–1112. https://doi.org/10.1002/asi.24460

Lightfoot, E. A. (2016). The persistence of open access electronic journals. *New Library World*, *117*(11/12), 746–55. https://doi.org/10.1108/NLW-08-2016-0056

Meddings, C. (2011). Digital preservation: The library perspective. *The Serials Librarian*, *60*(1–4), 55–60. https://doi.org/10.1080/0361526X.2011.556437

Mering, M. (2015). Preserving electronic scholarship for the future: An overview of LOCKSS, CLOCKSS, Portico, CHORUS, and The Keepers Registry. *Serials Review*, *41*(4), 260–265. https://doi.org/10.1080/00987913.2015.1099397

Moulaison, H. L., & Million, A. J. (2015). E-publishing in libraries: The [digital] preservation imperative. *OCLC Systems & Services: International Digital Library Perspectives*, *31*(2), 87–98. https://doi.org/10.1108/OCLC-02-2014-0009

Pendergrass, K. L., Sampson, W., Walsh, T., & Alagna, L. (2019). Toward environmentally sustainable digital preservation. *The American Archivist*, *82*(1), 165–206. https://doi.org/10.17723/0360-9081-82.1.165

Project JASPER. (2020). Directory of Open Access Journals. https://doaj.org/preservation/

Regan, S. (2016). Strategies for expanding e-journal preservation. *The Serials Librarian*, *70*(1–4), 89–99. https://doi.org/10.1080/0361526X.2016.1144159

Reich, V. (2006). Follow the money! *Serials Review*, *32*(2), 68–69. https://doi.org/10.1016/j.serrev.2006.03.008

Reich, V., & Rosenthal, D. (2009). Distributed digital preservation: Private LOCKSS networks as business, social, and technical frameworks. *Library Trends*, *57*(3), 461–475. https://doi.org/10.1353/lib.0.0047

Rieger, O. Y., & Wolven, R. (2011 December 8). *Preservation status of e-resources: A potential crisis in electronic journal preservation*. CNI: Coalition for Networked Information (blog). https://www.cni.org/topics/digital-preservation/preservation-status-of-eresources

Rosenthal, D. S. H. (2007 April 29). *Format obsolescence: Scenarios*. DSHR's Blog (blog). http://blog.dshr.org/2007/04/format-obsolescence-scenarios.html

Rosenthal, D. S. H. (2010). Format obsolescence: Assessing the threat and the defenses. *Library Hi Tech*, *28*(2), 195–210. https://doi.org/10.1108/07378831011047613

Rosenthal, D. (2020 September 17). *Don't say we didn't warn you*. DSHR's Blog (blog). https://blog.dshr.org/2020/09/dont-say-we-didnt-warn-you.html

Salo, D. (2020). Is there a text in these data? The digital humanities and preserving the evidence. In M. P. Eve & J. Gray (Eds.), *Reassembling scholarly communications: Histories, infrastructures, and global politics of open access* (pp. 215–228). The MIT Press. https://direct.mit.edu/books/book/4933/chapter/625170/Is-There-a-Text-in-These-Data-The-Digital

Seadle, M. (2011). Archiving in the networked world: By the numbers. *Library Hi Tech*, *29*(1), 189–197. https://doi.org/10.1108/07378831111117001

Sprout, B., & Jordan, M. (2018). Distributed digital preservation: Preserving open journal systems content in the PKP PN. *Digital Library Perspectives*, *34*(4), 246–261. https://doi.org/10.1108/DLP-11-2017-0043

Tkaczyk, D., Datta, E., & Crossref. (2023 February 14). Sampling Framework. GitLab. https://gitlab.com/crossref/sampling-framework

Van de Sompel, H., Rosenthal, D. S. H., & Nelson, M. L. (2016). Web infrastructure to support e-journal preservation (and more). *arXiv*. https://doi.org/10.48550/arXiv.1605.06154

Waters, D. J. (2005). *Urgent action needed to preserve scholarly electronic journals*. Association of Research Libraries. https://www.arl.org/resources/urgent-action-needed-to-preserve-scholarly-electronic-journals/

Waters, D. J. (2007). Preserving the Knowledge Commons. In C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: From theory to practice* (pp. 145–167). The MIT Press.

Wittenberg, K., Glasser, S., Kirchhoff, A., Morrissey, S., & Orphan, S. (2018). Challenges and opportunities in the evolving digital preservation landscape: Reflections from Portico. *Insights: the UKSG journal*, *31*(0), 28. https://doi.org/10.1629/uksg.421