Challenges and opportunities in the evolving digital preservation landscape: reflections from Portico

There has been tremendous growth in the amount of digital content created by libraries, publishers, cultural institutions and the general public. While there are great benefits to having content available in digital form, digital objects can be extremely short-lived unless proper attention is paid to preservation. Reflecting on our experience with the digital preservation service Portico, we provide background on Portico's history and evolving practice of sustainable preservation of the digital artifacts of scholarly communications. We also provide an overview of the digital preservation landscape as we see it now, with some thoughts on current requirements for preservation, and thoughts on the opportunities and challenges that lie ahead.

Keywords

Digital preservation; archives; electronic publishing; digital repositories; national libraries; text and data mining; open access; scholarly record

KATE WITTENBERG
Managing Director, Portico
ITHAKA

SARAH GLASSER
Director of Product Marketing, Portico
ITHAKA

AMY KIRCHHOFF
Archive Service Product Manager, Portico
ITHAKA

SHEILA MORRISSEY
Senior Researcher, Portico
ITHAKA

STEPHANIE ORPHAN
Director of Publisher Relations, Portico
ITHAKA

Introduction

In the last three decades there has been tremendous growth in the amount of digital content created by libraries, publishers, cultural institutions and the general public. There are great benefits to having content available in digital form. However, unlike print objects – which, when they have been printed on acid-free paper and are held in reasonable conditions, can last for many decades with only minimal attention – digital objects can be extremely short-lived unless proper attention is paid to preservation. Long-term preservation of this digital content is a key concern for the scholarly community as we continue to make the print-to-digital transition.

'digital objects can be extremely short-lived unless proper attention is paid to preservation'

Many organizations provide digital preservation, including governmental agencies, not-for-profits with preservation services, libraries and commercial entities interested in preserving their content for future research. It is over 15 years since the creation of the Digital Preservation Coalition and the early phase of community-supported services such as Portico and LOCKSS, and during this time we have learned a great deal about the challenges that exist in providing reliable, scalable and sustainable long-term preservation and, when needed, access to preserved content.

Reflecting on this experience and the continuing need for resources and attention to digital preservation, we provide here the background on Portico's history and evolving practice of sustainable preservation of the digital artifacts of scholarly communication. We also provide an overview of the digital preservation landscape as we see it now, along with our thoughts on the current requirements, opportunities, and challenges that lie ahead.

Looking back: Portico case study

In 2003 libraries and publishers alike were reeling from the RoweCom/Faxon bankruptcy. The RLG/OCLC report, *Trusted Digital Repositories: Attributes and Responsibilities*,[1] was less than a year old. By 2005, the year that marked the official launch of Portico, the report from the Andrew W Mellon Foundation, *Urgent Action Needed to Preserve Scholarly Electronic Journals*,[2] was released with wide support from the academic library community. Third-party preservation services such as Portico emerged in this context, in response to the needs articulated by the academic library community, whose increasing adoption of e-journals and other online resources provided enhanced ease of access to content, but required giving up traditional physical control over maintaining it.

Evolving service model: e-books, e-journals, digitized collections

The scope of Portico's work for the first several years of its existence was the preservation of e-journal content, with the ultimate goal of providing the academic community with access should that content no longer be available online through the publisher or a successor (in Portico parlance, a 'trigger event'). The cost of doing the important work of preservation was shared by the academic library and scholarly publishing communities, thereby ensuring that a broad range of content could be preserved, and that the service would be sustainable over the long term. By 2009, Portico began preserving e-book content, also on a community model, and had also launched a service specific to digitized historic collections, which is exclusively publisher supported.

Evolving community support

In 2006 Portico was supported by 27 publishers and 245 libraries. Currently, there are 518 publishers (representing over 2,000 learned societies and associations) participating in Portico. Committed content from these publishers comprises 28,468 e-journal titles, 1,231,039 e-book titles and 187 digitized collections. Library support has also grown steadily over the last decade. Today, more than 1,000 libraries around the world are Portico participants.

The original drivers for third-party preservation services still exist. Indeed, they have, if anything, intensified with the dramatic increase in both the breadth and depth of scholarly content available online. Libraries are increasingly opting for online-only access to journals and books, foregoing print altogether. Many resources are now available only in an online format from content providers. The shift from digital content as a secondary access mode for print, to digital content as the version of record and primary access mode, makes more acute the necessity for robust, third-party preservation and the assurance of future access that it provides.

'the necessity for robust, third-party preservation'

Ultimately, preservation services exist to ensure that scholarly content made available online today will always be available in the future. Library support for, and participation in, third-party services is predicated on the promise of ongoing usability and future access when necessary. Portico and other preservation services fulfill this promise through trigger events – the mechanism by which content is made accessible when, as mentioned, it becomes unavailable from the original publisher or successor.

Trigger events occur for a number of reasons, ranging from a publisher going out of business, to a publisher being acquired but without its discontinued titles being migrated to the acquiring publisher's platform, to a publisher actively deciding no longer to host a publication due to lack of financial support. In Portico's original service model, triggered content was made available only to libraries participating in Portico (regardless of whether they had ever subscribed to the triggered title). Over time, as open access (OA) moved from a peripheral publishing model to the mainstream, Portico changed its policy to meet the community's expectation that open content would remain open, and we now make it

possible for OA publishers to ensure that their content will remain open when triggered.

To date, Portico has triggered content from 46 e-journals, 25 of them OA. We have also triggered a reference work and an abstracts database. In addition to making content available as the result of trigger events, Portico, by agreement with some of its publishers, also makes content available as a post-cancellation/perpetual access provider for a large number of the titles it preserves. As of December 2017, content from 424 journals and 404 books was made available to 251 institutions through post-cancellation/perpetual access.

'we now make it possible for OA publishers to ensure that their content will remain open when triggered'

While all of our archive content has grown, e-journals continue to be our core service, and although many journals are preserved in Portico, there are many more titles in existence. A large number of them come from smaller publishers (defined as those that publish ten or fewer titles). These journals are arguably most at risk of disappearing if they are not preserved. However, it is a bigger challenge to get these journals into services like Portico. First, despite the low fees for these smaller publishers (fees being based on revenue), even a nominal fee can be a financial barrier for a publisher with low revenues and limited institutional support. Second, some are just so small that organizational and resource constraints make it difficult for them to participate. For example, many lack established processes (both technical and administrative) to authorize and manage content deposits for preservation. Portico has been able to make inroads in this area by developing tools to support smaller publishers (such as a content export plug-in for the Open Journal Systems), and by developing more flexible mechanisms for obtaining metadata and content deposits. At present Portico preserves content from more than 200 small publishers. While this represents good progress, there is still substantial work to be done in this area; as a point of reference, we know that there are at least 7,000 publishers listed in Crossref that do not have content preserved in Portico or other preservation services, and estimate that 4,700 of those are small publishers.

Evolving service model: national library collaboration

In order to leverage preservation efforts for scholarly materials, it has proved useful for organizations to explore ways in which they can co-ordinate efforts and collaborate. One such instance of collaboration is the relationship between Portico and the British Library.

In April 2013 the UK implemented legislation requiring the legal deposit of digital content published in the UK, with the British Library acting as the legal deposit library and co-ordinator of the program. The legislation is broad, encompassing 'e-books, e-journals and other types of electronic publication, plus other material that is made available to the public in the UK on handheld media such as CD-ROMs and microfilm, on the web (including websites) and by download from a website'.[3] One segment of the UK legal deposit legislation covers e-journals, many of which were already being deposited and preserved in Portico. In 2013 Portico entered into an agreement with the British Library to provide an e-journal processing service that supports the Library's legal deposit requirements. This relationship allows the British Library to leverage the existing Portico technical infrastructure and staff expertise. Portico is already working with varied content streams from publishers, standardizing article metadata, and reorganizing articles and issues into a package that fits our preservation content model. This 'normalized' content is then exported from Portico systems to the British Library, which provides them with a single stream of content to manage.

As of May 2018, Portico is delivering current e-journal content from 21 publishers to the British Library, comprising nearly 7,000 journals, with over 3.7 million articles delivered. The British Library ingests the content from Portico into its own preservation system and then pushes the content out to the deposit libraries throughout the UK and Ireland, where the e-journal articles are delivered to library patrons in reading rooms.

As content that requires preservation increases in both size and complexity, these types of

collaborative relationships will be key to cost-effectively meeting the preservation needs of the community.

Looking around: the scholarly communications landscape

Given the needs and challenges that motivated the creation of Portico 15 years ago, what are some of the challenges and opportunities that exist in the current and future digital preservation landscape?

Challenge: economics

A perennial challenge involves economics, particularly for higher education institutions concerned with preserving scholarly journals, monographs and special collections. There is a continued downward trend in public money to support higher education, along with increasing politicization around discussions of education policy and practice. The ripple effects from this trend affect research libraries, which have been required to work with diminished resources, and are thus increasingly constrained in their ability to support preservation services. For example, the majority of university libraries do not participate in any digital preservation service. Downward pressure on library budgets has meant that many institutions must focus on putting funds towards content acquisition rather than preservation, while smaller libraries may view preservation as outside their institutional mandate.

'collaborative relationships will be key to cost-effectively meeting the preservation needs of the community'

Challenge: new models of publication and reuse

Academic and research libraries have themselves increasingly played the role of publishers – of 'e-journals, conference proceedings, technical reports, and database-driven websites'.[4] More challenging, from a preservation point of view, is the role of libraries as custodians of what might be called 'locally created content': institutional repositories and other systems that contain specialized collections, teaching materials, preprints and other objects. While larger research and national libraries generally budget for preservation services (for both internal and external content), many smaller institutions have rare and unique special collections that require preservation. Without sufficient funding for these smaller libraries and cultural institutions, these special collections will not be preserved, which will put this valuable content at risk over the long term.

Traditional publishers are finding that they must respond to increased pressure for open access to the artifacts of scholarly research. The push for OA has been markedly increased by the 2001 Budapest Open Access initiative,[5] by the 2003 Bethesda Statement[6] and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities,[7] by the publication in June 2012 of the Finch Report,[8] and by the February 2013 Office of Science and Technology Policy (OSTP) memo, *Expanding Public Access to the Results of Federally Funded Research*.[9] It should be noted, however, that priorities for preservation tend to organize around content that libraries are paying to access, although OA content is also in need of preservation.

'priorities for preservation tend to organize around content that libraries are paying to access'

Partly as a result of initiatives to expand public access to research results, such as the OSTP memo and the FAIR Guiding Principles for scientific data management and stewardship,[10] there is also an increased focus on the capture, preservation and sharing of data sets and other research artifacts, and the creation of data management plans as part of all grant proposals in order to accomplish this. Preservation, which Portico defines as providing usability, authenticity, discoverability, and accessibility of content over the long term, necessarily has to engage with issues of data availability for reuse, including making large or complex data sets intelligible via visualizations and other representations, and of ensuring the discoverability of non-textual artifacts such as images, video, engineering and architectural models, geographic information systems and other specialized data formats.

Challenge: digital object complexity

Indeed, beyond data sets and non-textual artifacts, the increasing complexity of digital content is also becoming a challenge in its own right. As described in OCLC's 2014 report, *The Evolving Scholarly Record*,[11] the content that comprises the scholarly record has become, relative to its previous written, printed and even early digital forms, both more dynamic and less 'bounded'. Formerly, even a digital artifact of the scholarly record comprised a more or less discrete object, such as a journal article or book, often encapsulated in a single file. Increasingly, such an artifact is likely to be what we have come to call a 'distributed scholarly compound object'. Its various components can reside in more than one repository (including in preprint repositories) and have more than one version, mediated by complex hierarchies of continually changing hardware and software platforms. For such objects to comprise a scholarly record, and for that record to be preserved, preservation archives need to create, capture and maintain even more information about context and relationships than is provided in 'classic' bibliographic metadata. Such content also results in increasingly complex demands for discovery, access and rendering.

'the scholarly record has become … both more dynamic and less "bounded"'

Further complicating the challenges of stewardship are what Clifford Lynch calls 'algorithmically driven systems',[12] such as popular social media applications, which now also comprise nodes in the new complex network of scholarly communications.

How do we meet these challenges?

Opportunity: collaborations

A key response from the digital preservation community is collaboration – at scale and across discipline, professional and national boundaries. Portico itself is based on collaborative engagement by both libraries and publishers to ensure the long-term preservation of digitized and born-digital scholarly journals, books, reference materials and primary source collections. Portico both depends on, and contributes to, the international community of practice that is continually engaged in experimenting, defining standards and best practice, and developing toolsets and frameworks for preservation. Our work with the British Library is one example of how we have combined the work of two independent but like-minded institutions in the service of a collaborative partnership that benefits scholarship.

More generally, university and research libraries have taken collective action to meet the challenges of digital preservation. In the US, notable collaborative preservation initiatives include the Digital Preservation Network, CLOCKSS and the various Private LOCKSS Networks. Portico is also increasingly being approached by library consortia (such as the Swiss Consortium, the Deutsche Forschungsgemeinschaft in Germany and CAPES in Brazil) that are looking for national digital preservation strategies. Portico has additionally collaborated with partners in preservation research projects, including a recent two-year Alfred P Sloan Foundation-funded initiative, undertaken with the Johns Hopkins University Library Data Conservancy and IEEE, to develop the means to preserve the complex relationships among scholarly publications and their underlying data, thereby supporting the continual development of scholarly communication and digital publishing.[13]

Opportunity: advances in technology

The international community of preservation practice has also expanded beyond its center in libraries and archives, resulting in new perspectives, practices, tools and techniques. The need of the scientific community, for example, to ensure reproducibility of experimental results, both draws on and enriches the approaches of more 'traditional' preservation institutions. The participation of business, financial and governmental institutions in the Digital Preservation Coalition is an indicator of a broadening realization of the need to provide for the preservation of digital artifacts. Institutions such as VIAA in Belgium and Indiana University in the US in its Media Digitization and Preservation Initiative have undertaken creative new partnerships with private industry to develop solutions to the challenges of preserving audio-visual artifacts in legacy media and formats.

There have also been significant technical developments over the last ten years that are being embraced by preservation organizations (including Portico, the British Library, the Stanford Digital Repository and the UK National Archives) that have recently undertaken, or will soon undertake, the engineering of their next-generation preservation technology infrastructures. The palette of preservation architecture technologies has been expanded to include commercial and non-commercial cloud storage and server platforms (such as Amazon's AWS[14] and Microsoft's Azure[15]). Perhaps more important, many of the software libraries and APIs used by large-scale commercial application and platform service providers have been released as free and open source software, providing well-supported tools for horizontal scalability of both storage and servers (crucial for sustainably managing the growth of archives in size and complexity).

At the heart of the challenge of long-term access to digital artifacts is the continual and rapid obsolescing of the hardware and software that was originally employed for the creation and rendering of those artifacts. Another technology – emulation – provides an additional tool to help solve this problem of technical obsolescence. Emulation is the use of specialized software (and, sometimes, hardware) to enable a computer system to emulate, or behave like, another system – including the older hardware, operating systems and programs that provide access to older digital artifacts. Advances in emulation, along with research projects at the University of Freiburg[16] and Yale University,[17] have created new capabilities for employing emulation in the service of preservation.

'new capabilities for employing emulation in the service of preservation'

The large-scale development and open availability of natural language processing, machine learning and other text and data mining tools have the potential to automate metadata extraction and enrichment – both of which can be difficult and costly to do at scale. It remains an open question whether the use of these tools will result in the collection of 'good enough' bibliographic metadata in ways that might ease at least some of the work of preserving the long tail of scholarly journals. Supported by a recently announced grant from the Mellon Foundation,[18] the Internet Archive will be exploring the use of some of these tools in analyzing large-scale web harvests, with the intent of decreasing the burden on the library community to identify and prioritize which of the many long-tail publications should receive priority and attention for preservation.

Looking ahead

The challenge of digital preservation continues to grow, and because the artifacts collected and preserved are mediated by software, providing meaningful access to this content will be increasingly complex. As noted above, the rise of interdisciplinary and collaborative approaches to research, often spanning multiple disciplines and integrating text, data and audio-visual material, has enabled new forms of scholarly output. What can be done towards making at least some of these truly 'born-digital' artifacts also 'born-preservable'?

'The current infrastructure … may be insufficient to support new scholarly activities'

The current infrastructure that exists may be insufficient to support new scholarly activities along a spectrum from informal collaboration to stable, catalogued versions of work that are suitable for publication and long-term preservation. There is a need for the community to understand what will be required to support these new forms of content as they move through the scholarly communications process. How should we design the infrastructure needed to support this work? How do we preserve academic output that emerges from the less formal modes of research and discussion that scholars undertake outside the traditional publishing cycle? The willingness and ability of the community to provide the guidance and support required by scholars will play an important role in enabling the future creation, dissemination and preservation of this work.

What does Portico's past suggest to us, as we contemplate this future? As the number of both libraries and publishers participating in preservation services has grown, these communities have seen first hand that the investment in what was originally seen as an

experiment has paid off in a robust, large-scale, economically sustainable third-party preservation service. What we have learned at Portico, in the course of what we consider to be an ongoing experiment, is that the nature of the content we preserve, the technology that supports the service, the business model and characteristics of the service, the networks of partnerships and our relationships with those partners, have all evolved, and must and will continue to do so. While our mission remains the same as it was at Portico's inception, the ability to keep up with the pace of change requires agility, flexibility and a deep understanding of the evolving scholarly ecosystem.

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'Abbreviations and Acronyms' link at the top of the page it directs you to: **http://www.uksg.org/publications#aa**

Competing interests

The authors have declared no competing interests.

References

1. *Trusted Digital Repositories: Attributes and Responsibilities*, RLG-OCLC Report, 2002: **https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf** (accessed 23 May 2018).
2. Waters D, *Urgent Action Needed to Preserve Scholarly Electronic Journals*, Andrew W Mellon Foundation, 2005: **https://old.diglib.org/pubs/waters051015.pdf** (accessed 23 May 2018).
3. The British Library legal deposit for websites and electronic publications: **http://www.bl.uk/aboutus/legaldeposit/websites/** (accessed 23 May 2018).
4. Skinner K, Lippincott S, Speer J and Walters T, Library-as-Publisher: Capacity Building for the Library Publishing Subfield, *Journal of Electronic Publishing,* 2014, 17(2); DOI: **https://doi.org/10.3998/3336451.0017.207** (accessed 23 May 2018).
5. 2001 Budapest Open Access Initiative: **http://www.budapestopenaccessinitiative.org/** (accessed 23 May 2018).
6. 2003 Bethesda Statement on Open Access Publishing: **http://legacy.earlham.edu/~peters/fos/bethesda.htm** (accessed 23 May 2018).
7. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities **https://openaccess.mpg.de/Berlin-Declaration** (accessed 23 May 2018).
8. Finch J, *Accessibility, sustainability, excellence: how to expand access to research publications,* Report of the Working Group on Expanding Access to Published Research Findings, 2012: **https://www.acu.ac.uk/research-information-network/finch-report-final** (accessed 7 June 2018).
9. Holdren J, *Expanding Public Access to the Results of Federally Funded Research*, Office of Science and Technology Policy, 2013: **https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf** (accessed 23 May 2018).
10. Wilkinson M D et al., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data,* 2016; DOI: **https://doi.org/10.1038/sdata.2016.18** (accessed 23 May 2018).
11. Lavoie B et al., *The Evolving Scholarly Record,* OCLC Research, 2014: **http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf** (accessed 23 May 2018).
12. Lynch C, Stewardship in the 'Age of Algorithms', *First Monday*, 2017, 22(12); DOI: **https://doi.org/10.5210/fm.v22i12.8097** (accessed 23 May 2018).
13. RMap Project: **http://rmap-project.info/** (accessed 23 May 2018).
14. Amazon AWS: **https://aws.amazon.com/** (accessed 25 April 2018).
15. Microsoft Azure: **https://azure.microsoft.com/en-us/?v=18.10** (accessed 23 May 2018).
16. bwFLA – Emulation as a Service, the University of Freiburg: **http://eaas.uni-freiburg.de/** (accessed 23 May 2018).
17. Pevner J, 5 March 2018, Yale announces software recovery project, Yale News: **https://yaledailynews.com/blog/2018/03/05/yale-announces-software-recovery-project/** (accessed 23 May 2018).
18. Jefferson, 5 March 2018, Andrew W Mellon Foundation Awards Grant to the Internet Archive for Long Tail Journal Preservation, Internet Archive Blogs: **https://blog.archive.org/2018/03/05/andrew-w-mellon-foundation-awards-grant-to-the-internet-archive-for-long-tail-journal-preservation/** (accessed 23 May 2018).

Corresponding author
Sarah Glasser
Director of Product Marketing, Portico
ITHAKA, US
E-mail: sarah.glasser@ithaka.org

To cite this article:
Wittenberg K, Glasser S, Kirchhoff A, Morrissey S and Orphan S, Challenges and opportunities in the evolving digital preservation landscape: reflections from Portico, *Insights*, 2018, 31: 28, 1–8; DOI: https://doi.org/10.1629/uksg.421

Published by UKSG in association with Ubiquity Press on 17 July 2018