Preserving Innovation: Ensuring the Future of Today's Scholarship

KAREN HANSON

Portico is a preservation service that works with libraries and publishers to ensure that scholarly content is available for future scholars. When it was formed in 2005, the initial workflows were built around a concept of a publication that had remained relatively unchanged for centuries. As books and journals moved to digital formats, they mostly simulated the bound print world, retaining and adapting many of its artifacts and processes. While traditional publications have not changed significantly and still play a central role in scholarly communication, scholars now also have the option to use an exploding variety of platforms and tools to share their work in new and sometimes complex forms. The result is that the relative uniformity of traditional electronic publications, which enables scalable preservation, can no longer be assumed. Those who seek to preserve scholarship are confronted with publications that incorporate an ever-expanding variety of embedded media formats and viewers, data visualizations, version management, complex interdependent networks of supporting materials such as software and data, reader-contributed content (annotations, comments), interactive features, and nonlinear forms of navigation. Preservation services that focus on scholarly content continue to evolve in order to support these changes, and new services have arisen to meet demand in specific areas, such as software preservation. But as content becomes more complex and prolific, it is challenging for preservation services to keep pace with increasingly diverse publication formats and preserve them at scale. If authors and their publishers do not plan for the longevity of their work, the most innovative scholarship today may lose the characteristics that make it unique and valuable in a matter of years rather than decades. Preservation copies, in turn, may be missing important components of the work.

It is in this context that NYU Libraries proposed a project that would bring together open access publishers concerned with the long-term survival of their most innovative projects with digital preservation services that specialize in scholarly publications. The project, titled Enhancing Services to Preserve New Forms of Scholarship, was funded by The Andrew W. Mellon Foundation and had two goals. The first was for the preservation services, Portico and CLOCKSS, to test the limits of current preservation tools

https://doi.org/10.3998/jep.1852

for preserving new forms of scholarship at scale. The second was to generate a set of guidelines for publishers that could be used to design publications that are more likely to be preservable and sustainable over the long term. It was also an opportunity to start a conversation between preservation services and publishers about ways to collaborate on the shared goal of perpetuating access to unique and often costly publications.

Five publishers—Michigan Publishing, Stanford University Press, University of Minnesota Press, University of British Columbia Press, and New York University Press—shared 20 innovative examples of enhanced digital scholarly publications to be analyzed for preservability. While analysis showed that it was *possible* to preserve many of the publications in some reasonable form, these methods took a lot of manual effort and often collapsed if the approach needed to scale up. The interwoven and networked nature of modern digital publications means that content from different platforms and sources can be presented seamlessly through a single user interface to form a publication. A YouTube video, for example, can be visually presented as part of an article even though the video and player are on the YouTube platform. In addition to creating technical challenges for preservation, this adds complexity to rights management because a preservation service may require an appropriate license to collect and preserve both the publisher and third-party content. Beyond that, the addition of reader-contributed content introduces ethical concerns, such as privacy and individual rights, further complicating the mechanisms for scaling the preservation process. Through this project, the preservation services identified and documented challenges and potential solutions for scaling preservation of this complex content. These findings contributed to the first draft of *Guidelines for Preserving New Forms of Scholarship* (Greenberg, Hanson, and Verhoff 2021a), a document that describes changes to the creation process that can improve the longevity of complex works by facilitating long-term preservation and often also shorter term sustainability.

Findings

Drawing from the research described above, this section summarizes some of the patterns identified and aligns them with ideas for how publishers and preservation services can evolve together to accomplish scalable preservation of new forms of scholarship.

The complexity of "supplements"

Without the constraints of the bound form and with the falling cost of storage, sharing of additional resources to accompany a text has become commonplace. This has

been buoyed by funder mandates on data sharing and cultural changes that call for researchers to show the evidence underlying their work. In new forms of scholarship that integrate different kinds of materials, what were once called "supplements" may now be referenced and embedded throughout the publication in addition to sitting alongside them. This prompted several of the platforms involved in this research to call these materials publication "resources" rather than "supplements." As an integral part of the publication, these resources are vital to preserve in order to fully understand the intellectual contribution of the scholarship.

More than half of the publications analyzed had hundreds of resources. Resource file formats ranged from PDFs, audio, and video to executable programs and full databases. In three platforms, each of these resources included rich descriptive metadata, a dedicated landing page, and in some cases a unique persistent identifier making it possible to cite them directly. One third of the publications had resources of more than a gigabyte (GB) of disk space—much larger than a typical PDF publication.

For traditional publications in Portico, where supplements are supplied for preservation, they are packaged and archived with the ebook or ejournal and mostly have little or no metadata outside of the notes that are within the publication text. In these new forms of publication, Portico may need to reflect this rich expression of the associated resources with independent landing pages, improved support for a wider variety of file formats, and sometimes DOIs that can resolve to the archived version if the publisher copy is no longer available. This is all required while also ensuring that the resources remain linked to the text so that they can be presented as part of it through Portico's access platform. Increasingly, publications are a web of connected resources in which the text itself connects many pieces rather than a bound format that can be easily contained as a distinct object.

As preservation services work to support this evolving concept of supplements, publishers can increase the success of this effort in a number of ways. These are laid out in detail in the *Guidelines* but include the following key elements: supplying rich structured metadata for all resources; using common, non-proprietary file formats (these are generally easier to preserve as they have tools and paths supported by the community); and ensuring that the rights for the resources have been secured and that these permissions are expressed in the metadata.

Living documents

Traditional publishing has strict schedules and workflows that lead up to a publication date, after which the publication does not change except through established and formal channels such as addendums, editions, and retraction notices. For new

forms of scholarship, the concept of "version of record" can be elusive and may need to be considered case by case. Some current platforms enable version updates after initial publication without a formal addendum or change to the DOI. Others support annotations and comments from readers that appear on the publication and, for some publishers, are considered an important aspect of the published work. Other publications exist in a perpetual draft state, where they are designed to iteratively change to incorporate feedback or new data, never reaching a final, official publication date.

This kind of *content drift* is expected on the web (Jones et al. 2016), and website archivists have come up with some strategies and standards to preserve and provide access to different versions of web pages as they shift over time. For scholarly content this drift can complicate citations, as persistent identifiers may point to evolving content. For preservation services such as Portico that were initially modeled on a much clearer concept of versioning, it becomes a challenge to reconcile the difference between one version and another and to ensure that new versions are recorded appropriately. The addition of user comments, for example, introduces not only versioning challenges but also questions about the ownership of those comments and whether a preservation service has a right to preserve them. For some content, these questions can extend to ethical and individual rights issues such as the "right to be forgotten."

Living documents may warrant an addendum to the usual conversation that occurs between the publisher and Portico when planning the preservation workflow. To support preservation, publishers may need to define which version or versions should be preserved. The preservation system will need to recognize when something has changed using agreed-upon criteria that will support making the necessary distinctions through the descriptive metadata. For user-contributed content, if a publisher believes it is important to preserve, there will need to be clarity about whether there are rights for a third party in order to do so. Changes to, or negotiations regarding, terms of use/service may be necessary to support preservation of this content.

Embedded resources

One of the most frequent features of new forms of scholarship is the embedding of a variety of resources that would typically not be found in traditional publications. Examples include audio, video, and complex data visualizations that are seamlessly integrated into the body of the text, just as figure graphics are embedded in traditional publications. The EPUB format, and web pages generally, support simple methods for embedding audiovisual material using HTML tags for image, video, and audio. If used as intended, these can be managed by preservation workflows. In the samples analyzed

during this research, however, almost all material that was not text or image was embedded into publications using an inline frame, or *iframe*. This is an HTML feature that enables the seamless visual embedding of the content of one web page into the content of another. Embedding YouTube videos into an article, for example, is done using an iframe.

A multitude of preservation challenges result from the use of iframes. The content presented in them may not be part of the publishing platform or even controlled by the publisher. This makes the content of iframes less likely to be included in platform exports and vulnerable to *link rot*, where links on the web tend to stop working over time. Some embedded resources lack a formal caption or citation within the text. If the link stops working, this can result in a publication that displays an empty box saying "page not found" and no explanation to the reader of what was once there. This is a threat for not only the preservation copy but also the copy presented on the publisher's website. Broken iframes were seen in several examples during the project. The uncertainty of what an iframe might contain—it can contain any public web page—is also challenging. It may be unclear whether the preservation service has the appropriate rights to copy the iframe content if it is not a part of the publisher's platform. Even if preservation services can identify which web pages they are permitted to archive, the quality of automated web page preservation varies widely.

Portico is exploring new ways to identify which resources should be included in the archived copy and how to ensure they are embedded at the appropriate position in the publication. This includes adding web page archiving tools to some preservation workflows. To help avoid omission of resources that are integral to the work, preservation services may need support from publishers and their platform designers. Workflows that standardize the expression of common embedded features within the publishing platform could be leveraged to design compatible workflows for preservation. Where possible, publishers could ensure they have a copy of all embedded resources, whether managed by the publisher platform or not, and the appropriate rights to preserve them. These copies could be included in preservation packages and also provide a useful backup for publishers in the case that an embedded resource becomes unavailable on the web even before the publication reaches an archive. Where copies or rights cannot be obtained, meaningful captions that include a description and an original link could help future readers find the content even if it is no longer connected to the publication. In some cases, publishers may wish to participate in a service that allows for static captures of web pages that are displayed or linked in the publication so that the content is preserved before it has a chance to change or disappear. An example of this type of service is Save Page Now, a free service by the Internet Archive that allows users to generate a snapshot of a web page and link to it.

The value of an experience

For some new forms of scholarship, enormous effort has gone into creating a specific presentation or *experience* of the work. These publications are among the greatest challenges for scalable preservation, as they significantly narrow the options for preserving the work. In essence, the work is bound to a particular technology stack or may require an unsustainable amount of development to preserve the experience on modern technology stacks through time. An example of an experience-rich monograph that was analyzed during this research is *As I Remember It: Teachings (ʔəms taʔaw) from the Life of a Sliammon Elder* (Paul et al. 2019). The popup "Protocol for being a respectful guest" that opens when visiting the site, the nonlinear navigation options, and the integrated audiovisual elements throughout the publication were among the dynamic elements defined by the publisher as being vital to the work.

For Portico, the primary mode of preservation has been to separate the intellectual components of the publication from the platform and arrange them into a standard package that can be updated and adapted through time to work with modern technology. If the experience is an important aspect of the publication that cannot be easily separated from the platform, then preservation approaches that record this experience should be considered.

One commonly used option for preserving the experience of a website is a web crawler, in which a tool visits and records a copy of the web pages from the outside. For websites that are compatible with this approach, the process can be automated with little configuration. Many websites, however, have features that require the crawler to be customized in order to record them. Some websites, such as those whose content can only be discovered via a search bar, cannot be preserved using this method. Another option is to recreate the website's server on a virtual machine and then preserve that virtual machine. This approach is rarely used for website preservation as it depends on access to the resources needed to recreate the server (code, data, software, licenses, documentation, expertise, etc.) and also whether the website can be configured to function without access to URLs outside of the web server. If the website centers on a visualization hosted by ArcGIS Online, for example, it will only work for as long as that visualization is available. There may also be some uncertainty about the technology and expertise required to run the server in the long term, but this is evolving as the tools and infrastructure to support this approach are making progress through efforts such as Emulation as a Service Infrastructure (EaaSI).

For web publications that integrate a lot of dynamic features and technologies, it can be difficult to apply website archiving methods and maintain high quality preservation at scale. Each of the web platforms analyzed during this research required several weeks of effort to create a web archiving process that met the preservation

requirements. Even with configuration tailored to the platform, unanticipated variations between publications and features that cannot be preserved using these methods can lead to an incomplete archival copy. Because the publications in this research were complex and spanned many web pages, the web archived content also required significantly more disk space compared to the exported content, which may have implications for the preservation cost at scale. If these challenges can be navigated, however, these approaches can be highly effective for platforms that favor them and may be vital to ensuring that the most innovative publications created today can be experienced in the future.

To respond to the need to preserve the experience for some publications, Portico has initiated a web archiving pilot project that uses a crawler and will continue to evaluate scalable options for preserving websites that cannot be crawled. While preservation services will always attempt to evolve and work with the content as presented, one way to improve the chance that these experience-focused approaches will be successful at high quality and scale is for publishers and preservation services to work together to ensure the platforms will favor these techniques. Numerous suggestions for how to make this possible are documented in the *Guidelines*. For web publications, publishers can support this work by following best practices that make their websites easier to archive from the outside, such as introducing sitemaps or other mechanisms to guide website crawlers to content that should be preserved. These practices offer the additional benefit of making the websites more discoverable by search engines. For publications that are data driven and cannot be made web crawler friendly, publishers may be asked to share source code, data, and resources that allow for the re-creation of the platform in order to help prevent loss. In general, using open source, non-proprietary technologies and avoiding use of complicated or obscure technologies where simple, broadly adopted technologies would suffice will also support this approach.

Conclusions

The previous section laid out ideas for how publishers and preservation services can work together to tackle different kinds of challenges. These are addressed in more detail through the *Guidelines* that resulted from this initial project. In addition, a full report describing the project is available (Greenberg, Hanson, and Verhoff 2021b). A new project, also funded by The Andrew W. Mellon Foundation, will further test this process. Embedding Preservability will embed preservation specialists into publisher workflows to help test and implement relevant *Guidelines*. The team of preservation specialists will also work with the corresponding platform developers to see if the recommendations described in the *Guidelines* might be built into their systems. The preservation services

meanwhile will work to evolve in the directions required to support these new forms of scholarship and test their ability to preserve the publications they have played a part in developing.

While the *Guidelines* are offered as a tool for publishers who may not have access to a preservation expert, some partners noted a potential emerging need for a closer relationship with preservation services, or even permanent staff members who are partly dedicated to ensuring that even the most innovative publications meet some basic requirements for sustainability and preservation. Stanford University Press has a position whose role includes working with authors and preparing each work for preservation. Michigan Publishing's Fulcrum platform was designed in collaboration with library preservation experts. University of Minnesota Press engaged with Portico early in their process of building an export tool for preservation into their Manifold platform. There is also some overlap between requirements for preservation and other incentives and standards. Publishers may meet multiple criteria by following standards for metadata, accessibility, research reproducibility, and search engine optimization and by working to alleviate longer term financial burdens and loss through the creation of publications that are easy to maintain and do not require ongoing communication with the author. New consulting roles for preservation services may emerge to support this work.

Every instance in which publishers and preservation services can collaborate to build tools and methods that can plug into platforms or be reused by others will contribute to an information infrastructure that favors the longevity of scholarship. Rather than asking publishers to stifle the innovation and creativity of their authors to simplify the preservation task, we are looking to collaborate and innovate to develop approaches to creating new forms of scholarship that will be available to future scholars.

References

European Union. 2018. "Art. 17 GDPR—Right to Erasure ('Right to Be Forgotten')." Accessed November 14, 2021. https://gdpr.eu/article-17-right-to-be-forgotten/.

Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. 2021a. *Guidelines for Preserving New Forms of Scholarship*. New York University, September. https://doi.org/10.33682/221c-b2xj.

Greenberg, Jonathan, Karen Hanson, and Deb Verhoff. 2021b. *Report on Enhancing Services to Preserve New Forms of Scholarship*. New York University, December. https://doi.org/10.33682/0dvh-dvr2.

Internet Archive. n.d. "Save Page Now." Wayback Machine. Accessed November 15, 2021. https://web.archive.org/save/.

Jones, Shawn M., Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover. 2016. "Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content." Edited by Neil R. Smalheiser. *PLoS ONE* 11, no. 12: e0167475. https://doi.org/10.1371/journal.pone.0167475.

New York University. 2019. "NYU Receives Major Grant from The Andrew W. Mellon Foundation; Collaborative Effort Aims to Meet the Challenge of Preserving New Forms of Digital Scholarship." News release, April 17. https://www.nyu.edu/about/news-publications/news/2019/april/nyu-receives-major-grant-from-the-andrew-w--mellon-foundation--c.html.

NYU Libraries. 2021. "The Andrew W. Mellon Foundation Awards NYU $502,400 for Libraries Project to Expand Capabilities for Preserving Digital Scholarship." *News and Stories*, August 4. https://guides.nyu.edu/blog/The-Andrew-W-Mellon-Foundation-Awards-NYU-502400-For-Libraries-Project-to-Expand-Capabilities-F.

Paul, Elsie, David McKenzie, Paige Raibmon, and Harmony Johnson. 2019. *As I Remember It: Teachings (ʔəms taʔaw) from the Life of a Sliammon Elder*. Vancouver: UBC Press. http://ravenspacepublishing.org/as-i-remember-it/.

Software Preservation Network. n.d. "Emulation-as-a-Service Infrastructure." Accessed November 15, 2021. https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/.