

# Ask the preservation experts at the inaugural NISO Plus conference

Stephanie Orphan<sup>a,\*</sup> and Craig Van Dyck<sup>b,\*</sup>

<sup>a</sup>*Director of Publisher Relations, Portico, New York, NY, USA*

<sup>b</sup>*Executive Director, CLOCKSS Archive, PO Box 1113, South Orange, NJ, USA*

**Abstract.** During the inaugural 2020 NISO Plus Conference that was held from February 23–25, 2020 in Baltimore, MD, several “Ask the Expert” sessions were scheduled so that attendees could have access to experienced information industry executives who could address questions on a variety of topics. This brief paper is based upon the session on the topic of the preservation of scholarly information featuring Stephanie Orphan, Director of Publisher Relations at Portico and Craig Van Dyck, Executive Director of the CLOCKSS Archive. The experts first fielded questions from the moderator, Wendy Queen, Director of Project Muse, who asked about preservation challenges, the role that scholars should play in having their works preserved, the preservation of outputs from thought-leadership conferences such as NISO Plus, standards across publishing that create a burden for preservation, etc. In the remaining time they answered questions from the audience.

Keywords: Preservation challenges, portico, CLOCKSS archive, digital scholarship

## 1. Introduction

On day two of February’s NISO Plus conference (the last face-to-face gathering that many of us attended!), Stephanie Orphan, Director of Publisher Relations at Portico (see: <https://www.portico.org/>), and Craig Van Dyck, Executive Director of the CLOCKSS Archive (see: <https://clockss.org/>) fielded questions from the audience during the Ask the Experts about Preservation session. CLOCKSS and Portico are third-party digital preservation services for scholarly content, supported by the academic library and scholarly publishing communities. E-journal preservation is the largest segment of both services, but other content, such as e-books and digitized primary source collections are also preserved. Increasingly there is a need to consider methods for new types of scholarship - and both Portico and CLOCKSS are part of a Mellon-funded grant project at New York University (Enhancing Services to Preserve *New Forms of Digital Scholarship*) [1] concerned with preservation of enhanced monographs. Although their approaches to preservation are different, Portico and CLOCKSS are aligned around the shared role of working with the community to ensure the long-term usability of scholarly content.

---

\*Corresponding authors: Stephanie Orphan. Email: [Stephanie.orphan@ithaka.org](mailto:Stephanie.orphan@ithaka.org); Craig Van Dyck. Email: [cvandyck@clockss.org](mailto:cvandyck@clockss.org).

## 2. Preservation challenges

The half-hour session was expertly moderated by Wendy Queen of Project Muse who got things started by asking what each expert sees as the biggest challenge of digital preservation. Both experts discussed the new challenges posed by emerging types of scholarship and the move towards dynamic content rather than traditional flat files. Portico and CLOCKSS are learning a lot through their participation in the *Preserving New Forms of Digital Scholarship* grant, which will help the organizations plan for working with nontraditional content types. Also mentioned was that there is still a lot of work to do around identifying and preserving content from the “long tail” of small journal publishers, and that there are decisions that need to be made around preserving preprints (Portico preserves these; CLOCKSS does not yet). A key takeaway is that preservation organizations need to figure out ways to evolve their scope while at the same time supporting their core business, and managing resources.

## 3. A role for scholars?

The question was raised as to whether there are things that scholars can do to ensure that their content is in a format that is easily preservable. The experts discussed that publishers and libraries typically act as proxies for authors and researchers in terms of ensuring that their content can be preserved. It would be difficult for services such as CLOCKSS and Portico to work directly with authors. However, guidelines for authors, particularly for those creating nontraditional content, are things that preservation agencies are looking to develop. In addition, as part of the previously mentioned grant project, Portico’s senior research developer is exploring preservation options for dynamic content on the Manifold and Fulcrum platforms. This involves experimenting with web crawling, from which it has been learned that moving conversations upstream so that the inputs meet specifications will greatly increase success. CLOCKSS routinely crawls publishers’ websites, and is working with Webrecorder (<https://webrecorder.net/>) to enhance the ability to capture more of the dynamic features.

## 4. Ensuring future usability

Several questions came from the audience related to file formats. An initial discussion around supplemental files expanded to a discussion around the fact that the more you adhere to standards, the better it is downstream. Some publishers require that authors upload their supplemental files, and they can range from traditional text files to programs and other discipline-specific formats. A member of the audience asked whether preservation agencies are ensuring that all of these disparate file formats will remain usable, particularly for example, files from a publisher’s back file. Will these formats still be usable fifty years into the future? It was acknowledged that fifty years is a long time, particularly where nontraditional file formats are concerned. There are preservation solutions for everything, but some approaches cost more than others, and there are scalability questions (for example, around emulation as a preservation solution). Supplemental materials are, however, considered to be part of the journal article and, therefore, should be preserved. Some publishers host supplemental files outside of their primary publishing platform and, therefore, cannot easily provide the content or allow it to be captured for preservation (Portico asks publishers to deposit content to them via FTP while CLOCKSS works with both FTP deposits and site crawling).

It was mentioned that Portico's migration-based approach to preservation is designed so that the service monitors file formats and is prepared to migrate file types as they become obsolete. For broadly used file types, however, - XML, PDFs, images - publishers are typically updating file types along the way. Portico maintains a format registry, so if formats can be identified, they will be fully preserved. File types that cannot be identified will receive byte-level preservation, which means that the files will be preserved but not migrated. The format registry is updated when file types are identified, so it is possible for something that started out byte-preserved to eventually be fully-preserved with a commitment to migration once the file format has been identified. CLOCKSS preserves the content in the forms that publishers make available and is prepared to migrate on a just-in-case basis.

The experts emphasized that the more standardized systems and formats are, the more likely it is that content can be properly preserved. Publishers are encouraged to review the outputs from NISO's committee on supplemental materials [2] and adhere to them, but there is more work that the community could do to ensure standardization in the handling of supplemental materials, which will increase the likelihood that these materials will be preserved.

Wendy commented that as an aggregator, she views preservation as a relay race - with a handoff that keeps happening. Project Muse has the publisher files that they deposit with Portico. She wondered if there is more that aggregators and platforms could communicate to publishers regarding files. The experts agreed that common formats and using standard identifiers are foundational. The large publishers and aggregators are responsible for the vast majority of content that flows to CLOCKSS and Portico for preservation, and they are following norms and providing consistently-formatted content. Because aggregators such as MUSE create standardized exports on behalf of the publishers, they are already solving what could be a problem for some publishers. More issues arise around content deposits from the long tail of very small publishers that do not always have the technical understanding or staffing to prepare and provide files or platform integration that are ideal for preservation.

## **5. Enquiring publishers want to know**

The session wrapped with a question to the audience about what sort of challenges publishers face related to preservation. One publisher expressed that with the move from print to online for the version of record, there was no longer necessarily front and back matter presented online. They wanted to know if there was a standard practice for archiving front and back matter and whether it was problematic when publishers don't post it. The response was that if a publisher does include front and back matter online, preservation agencies want to receive and preserve it, but it is not problematic if the files don't actually exist. If a journal were to no longer be hosted online, Portico and CLOCKSS want to be able to provide access to the same files users could get access to when the publisher hosted the content. Therefore, if files were never posted online, they don't need to be created or supplied specifically for preservation.

Another publisher wanted to know what the preservation point of view is around external content referred to via links within an article. The experts expressed that standard practice is to preserve the reference link, but not the third-party content. However, through recent research, preservation agencies now have a better understanding of the best methods to use to capture external content, creating a local copy to be preserved. Of course, there are rights issues surrounding third-party content, particularly content such as YouTube videos. CLOCKSS and Portico would assume that we do not have the rights to capture most external content, but determining that is part of a vetting process. It would be hard to check each item,

however, leaving content such as YouTube videos unpreserved except in cases where publishers provided explicit information around rights.

In a fitting wrap-up, there was a final question about whether or not papers and other documentation from conferences such as NISO Plus are preserved and the sentiment was expressed that these gatherings of thought leaders generate much valuable information. While the preservation agencies are not specifically working with conference organizers to preserve conference materials, when included as part of an organization's standard preservation agreement, conference proceedings are preserved. There is really no technical barrier to preserving such materials, but conference organizers would need to take initiative to investigate it and be willing to bear some cost.

### **About the Authors**

**Stephanie Orphan** is Director of Publisher Relations at Portico where she is responsible for maintaining and expanding publisher participation in the Portico preservation service to ensure ongoing growth and sustainability of the Portico archive. She brings a deep understanding of publishing platforms, metadata, and packaging formats to her role, as well as significant relationship-management experience. Stephanie holds a Master of Science in Library and Information Science from the University of Illinois at Urbana-Champaign. She is currently a member of the Board of Directors of the Open Access Scholarly Publishers Association (OASPA). E-mail: [Stephanie.orphan@ithaka.org](mailto:Stephanie.orphan@ithaka.org).

**Craig Van Dyck** has been the Executive Director of the CLOCKSS Archive since November 2015. Previously he was with Wiley for eighteen years as Vice President of Content Management; and with Springer New York for ten years, most recently as Senior Vice President and Chief Operating Officer.

Craig served as Chairman of the Association of American Publishers Enabling Technologies Committee from 1995–1998, and was instrumental in the development of the Digital Object Identifier (DOI) system of CrossRef. He represented Wiley on the Boards of Directors of the International DOI Foundation, CLOCKSS, ORCID, CrossRef, and the Society for Scholarly Publishing, and was a member of the Portico Advisory Committee. E-mail: [cvandyck@clockss.org](mailto:cvandyck@clockss.org).

### **References**

- [1] NYU Receives Major Grant from The Andrew W. Mellon Foundation; Collaborative Effort Aims to Meet the Challenge of Preserving New Forms of Digital Scholarship, Press Release, April 17, 2019, see: <https://www.nyu.edu/about/news-publications/news/2019/april/nyu-receives-major-grant-from-the-andrew-w-mellon-foundation-c.html#:~:text=New%20York%20University%20has%20received%20a%20grant%20of,preservation%20of%20complex%20new%20forms%20of%20digital%20scholarship>, accessed August 1, 2020.
- [2] Supplemental Journal Article Materials (NISO/NFAIS), see: <https://www.niso.org/standards-committees/supplemental>, accessed August 1, 2020.