

## **An Overview of Portico: An Electronic Archiving Service**

### **Abstract**

As scholars have become more reliant upon electronic scholarly journals, the long term preservation of these resources has become a growing and increasingly urgent need. In 2002, JSTOR launched a project which has now become Portico, a new, not-for-profit electronic archiving service established to address the scholarly community's critical need for a reliable means to preserve scholarly e-journals. Portico is focused on preserving the intellectual content of e-journals through source file normalization and format migration. To date seven publishers have agreed to deposit more than 3,000 journals in the archive, which will be sustained by contributions from publishers and libraries.

### **Introduction**

As scholars have become more reliant upon the convenience and availability of electronic versions of scholarly journals, the long term preservation of these resources has become a growing and increasingly urgent need. The Association of Research Libraries' recent survey of its membership found that on average, 31% of total library material expenditures are devoted to electronic resources.<sup>i</sup> Yet, even as libraries are expending significant funds on e-resources, many librarians are uncertain how future access to these resources will be assured. They are reluctant, therefore, to decrease expenditures associated with the receipt and storage of print journals until robust e-journal preservation arrangements are in place. Because library demand for print resources persists, scholarly publishers are similarly limited in their ability to decrease their investment in the production and distribution of print materials. To enable both libraries and publishers to move forward, the emergence of reliable electronic archiving solutions has become critical in order to facilitate the community's transition from reliance upon print resources to greater – and secure – reliance upon the electronic format. The endorsement in late 2005 by the Association of Research Libraries (ARL), the Association of College and Research Libraries (ACRL) and others of the [\*Urgent Action Needed to Preserve Scholarly Electronic Journals\*](#) statement reflects the urgency of the need for a robust archiving

solution and the recognition that now is the time for the library community to act in support of initiatives that will ensure enduring access to scholarly e-journals and to encourage publishers to support archival efforts as well.<sup>ii</sup>

Anticipating the need for robust preservation of electronic journals, in 2002, JSTOR launched a project which has now become Portico. Portico is a new, not-for-profit electronic archiving service established in order to address the scholarly community's critical and urgent need for a robust, reliable means to preserve electronic scholarly journals. Portico builds upon and advances JSTOR's efforts to provide a trusted and reliable community-based archive, and Portico works with JSTOR to significantly expand the preservation infrastructure developed on behalf of the scholarly community. JSTOR has provided initial support for Portico's development together with Ithaka, The Andrew W. Mellon Foundation, and the Library of Congress. Portico's mission is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers, and students. This paper provides a brief summary of the history of Portico, an overview of the Portico electronic archiving service, and concludes with a brief look at Portico's current status and future plans.

## **History**

Portico ([www.portico.org](http://www.portico.org)) began as the Electronic-Archiving Initiative, a project launched by [JSTOR](#) with a grant from [The Andrew W. Mellon Foundation](#) and intended to build upon the Foundation's seminal E-Journal archiving program. The charge of the project was to build an infrastructure and economic model able to sustain an electronic journal archive. Initially the focus of the project was designing and prototyping content handling and archival systems, crafting potential archive service models, testing possible models with libraries and publishers and a drafting business model able to support a long term archival effort. For more than two years, project staff worked on the development of technologies necessary to meet the project objectives and engaged in extensive discussions with publishers and libraries to craft an approach that balances the needs of both communities while researching what would be necessary to build a sustainable business model for the archive. During 2004 the project was transitioned

to [Ithaka](#), and efforts to hone the Portico archival service continued. These efforts involved wide-ranging discussions with a large and informal network of librarians from more than fifty academic institutions of all types and sizes and the engagement of ten publishers ranging from small scholarly societies, to a university press, to large commercial publishers who agreed to participate in the discovery phase of the project.<sup>iii</sup>

Our early work was informed by several assumptions regarding characteristics key to a long-term archival effort. These assumptions were drawn from the electronic journal preservation planning research program funded by The Mellon Foundation and from JSTOR's operational experience serving as a third-party archive since 1995. The May 2002 report, [Trusted Digital Repositories: Attributes and Responsibilities](#), produced by the Research Libraries Group and OCLC, also influenced the earliest stages of the project.<sup>iv</sup> Building from this collective base of experience, we assumed that a reliable long-term archive would require at least five basic elements: 1) an institutional mission with preservation at its core; 2) an economic model capable of sustaining the archival effort; 3) a robust and evolving technological infrastructure sufficient to meet the complexities of electronic resources; 4) relationships with libraries, the traditional preservation stewards, and 5) relationships with publishers, creators of the content that is to be preserved.

During the early months of the project we tested these assumptions as we worked extensively with the ten publishers who were part of our developmental phase with a wide base of libraries as well. The involved publishers submitted a wide selection of sample electronic journal data that enabled us to assess the technological challenges that an e-journal archive would need to surmount, and they also participated in discussions sharing their own perspective on the e-journal archiving challenge and the concerns and needs which they faced as content creators. These conversations were mirrored by those we initiated in the library community. We benefited from working with librarians at institutions ranging from very small liberal arts colleges to the largest research libraries and universities who generously shared their perspective on the electronic archiving challenge.

This community engagement and our work to build a prototype archive affirmed our beginning assumptions but also yielded several important lessons. First, it was clear that the preservation of electronic journals presented many significant technological challenges. Electronic-journals are produced in diverse formats, which may vary in format or data quality within a single title or publisher. Any long term preservation infrastructure - including software, hardware and staff with appropriate expertise - must respond to the challenges of this diversity. Second, the sheer scale of the digital material requiring preservation dictates that libraries and publishers build mechanisms for cooperative effort. No one place can do it all, and, because libraries in many cases lease rather than own electronic journals, the cooperation of the scholarly publishing community is essential. Securing this cooperation is complicated by the fact that preservation is not viewed as a mission-core activity for most publishers. Third, access to archived literature is a key issue for both publishers and libraries, but they view this issue from very different perspectives. Publishers are understandably eager to ensure that access to archived literature does not reduce the value of their current product offerings. Libraries are understandably focused on assuring that access to core literature for their campuses will be reliable and timely. Finally, all parties we engaged recognized the importance of building a robust economic model capable of sustaining a long term archive – perhaps the greatest challenge preservation raises.

Building from these assumptions and findings that emerged from our iterative and collaborative discussions with the community, a new service, now known as Portico, was shaped and launched in 2005.

### **Overview of the Portico Archival Service**

The Portico electronic archiving service is initially focused on the long-term preservation of electronic scholarly journals. The Portico archive, which is a centralized repository, is open to a scholarly publisher's complete list of journals, including those titles which may be published in electronic format only, or print and electronic formats, or which may have been “reborn” or digitized from print. Portico is focused on preserving the intellectual content of the electronic scholarly journal; we do not attempt to recreate or

preserve for the long term the exact look and feel of the journal or the publisher's website or delivery platform.

### ***Portico's Archival Approach***

Portico's archival approach for electronic journals is that of managed preservation focused on the publishers' e-journal source files. Source files are the electronic files containing graphics, text, or other material that comprise an electronic journal article, issue, or volume. Source files may differ from files presented online most typically by including additional information such as richer mark-up or higher quality graphics. Portico receives source files directly from the scholarly publishers who have agreed to contribute to the Portico archiving service. We take an extremely cautious approach and verify the integrity of the content submitted by publishers so that we can fulfill the preservation commitments that we make regarding the usability and sustainability of the content. For example, once a profile, a machine-readable set of publisher-specific instructions, is established for a given publisher's data stream, any change in file formats or naming conventions on the part of the publisher will raise a warning in our ingest system. All content is identified by an ISSN and is mapped to an archiving agreement; every file is virus-checked; file formats are identified and validated; and preservation metadata is created including provenance, technical metadata, and archival events. We also inspect a randomly selected portion of the content to assure that our processes are working correctly and that the content is what was expected.

Portico has adopted migration as its archival strategy, and therefore Portico subjects the publishers' source files to a systematic normalization process that facilitates preservation and future migration.<sup>v</sup> Consistent with Portico's conservative approach, when publishers' proprietary XML and SGML Document Type Definitions (DTDs) are normalized to the NLM Archive and Interchange DTD, the archival format which Portico and other preservation efforts have adopted, both the original and the normalized XML are retained in the archive.<sup>vi</sup> Retaining the original source files helps to ensure that if, for instance, a data processing error or data corruption is discovered, reprocessing can be done from the original source. The archive also includes all DTDs and schemas necessary to understand the archived content and accompanying

documentation, where available, and the entire archive can be recreated from the assets and the METS files without reference to any proprietary data structures other than human-readable XML Portico METS files. Portico takes responsibility for the long-term preservation and management of the archived materials.<sup>vii</sup>

### ***Accessing the Archive***

Portico's normalization efforts are focused on ensuring that content remains available and accessible into the future. As mentioned above, accessibility of the archived content is a key issue for both publishers and librarians. As noted in the *Urgent Action* statement referenced earlier, "preservation of electronic journals is a kind of insurance, and is not in and of itself a form of access."<sup>viii</sup> Portico recognizes that while access to e-journal literature today may not be a concern, librarians and their constituents do need to have assurance of future access, a theme echoed in the *Urgent Action* statement. To address this need, all libraries supporting the Portico archive have campus-wide access to archived content when specific trigger events occur, **and** when titles are no longer available from the publisher or other source. Trigger events include:

- A publisher stops operations; or
- A publisher ceases to publish a title; or
- A publisher no longer offers back issues; or
- Upon catastrophic and sustained failure of a publisher's delivery platform.

In addition to these trigger events, both publishers and libraries have recognized that in some cases, even after a library has terminated a license to an electronic resource, it may be necessary for that library to continue to have ongoing access. This is commonly known as "perpetual access" or post-cancellation access. A publisher may choose to extend perpetual access to a library and that access can be provided through the Portico archive, if the publisher desires. In addition, select librarians at participating libraries are granted password-controlled access to the archive for verification purposes. This verification access, which is granted to the entire archive, is not intended to be used as a replacement for commercial document

delivery services or to fulfill inter-library loan requests. Finally, all publishers participating in the archive have full access to their own content and any content for which a trigger event prevails.

### ***The Role of Publishers***

Archiving cannot be accomplished without the cooperation and participation of the content owners – publishers – but because archiving is not necessarily core to a publisher’s mission, it is important to keep barriers to participation in archival arrangements as low as possible. Portico does this by agreeing to receive source files in the format created by the publisher and placing the responsibility of content normalization in the hands of the archive. To participate in Portico, a publisher:

- Signs a non-exclusive archiving license that gives Portico the right to ingest, normalize, archive, and migrate the publisher’s content.
- Indicates whether Portico will serve as a perpetual access mechanism.
- Supplies electronic journal source files in a timely way.
- Makes an annual financial contribution.

### ***The Role of Libraries***

While preservation may not be mission critical for publishers, it is at the heart of the work of many libraries. And, as the *Urgent Action* statement makes clear, preservation of *electronic* resources is especially important for libraries. Although electronic resources have raised new preservation challenges, they also have generated new opportunities. Libraries preserved important print materials by holding content locally in substantial physical infrastructure, which was replicated in many library buildings at many academic institutions. Today, however, libraries can effect the preservation of electronic journals without incurring the burden and expense of creating many local instances of complex and costly technological infrastructure. By supporting cooperative efforts, such as Portico, libraries (and publishers) can collaboratively contribute toward a shared infrastructure which supports a mutually beneficial and

valued goal – the long-term, robust preservation of scholarly literature published in electronic form. To participate in the Portico archive, a library:

- Signs an archiving license agreement.
- Makes an annual support payment.
- Provides IP or other relevant information for user authentication purposes.

### ***Sustaining the Archive***

Financial support is critical to a long-term preservation effort of any kind. Two kinds of support are needed: funds for initial development of technological infrastructure and early operations, and funds that can support the operation of the archive over time. Portico has secured substantial grant support from JSTOR, Ithaka, the Library of Congress, and The Andrew W. Mellon Foundation to cover the costs of initial development. Efforts are now underway to begin to receive and archive e-journal content. Unlike in the for-profit sector, the “venture capital” funds made available for Portico’s start-up phase do not need to be repaid.

While Portico will not attempt to recover its initial funding, as it becomes operational, the organization will need to cover its ongoing operating costs from diversified funding sources; it cannot rely upon any single revenue stream. The chief beneficiaries of the archive - publishers and academic institutions – will provide the primary sources of funding; however, charitable foundations and government agencies will also be expected to provide support. Publishers are asked to make an annual contribution to the archive and share with other supporting publishers the ongoing costs to receive, normalize, store, and migrate journal content. Fees are based on publishers’ total journals revenues (subscription, advertising, licensing) and range from \$250 to \$75,000 per year.

Libraries are also asked to make an annual payment to support the ongoing work of the archive, including the addition of new content to the archive, maintenance and enhancement of the technological infrastructure, and format migrations as technology evolves. Library Annual Archive Support payments

are tiered and vary according to a library's self-reported total Library Materials Expenditure (LME), reflecting Portico's value in preserving a growing portion of a library collection. Fees range from \$1,500 to \$24,000 per annum. To encourage broad participation in Portico from the outset, from institutions of all types and sizes, Portico designates early participants (institutions who begin Portico support in 2006 and 2007) as "Portico Archive Founders", and recognizes their early support of this important initiative by providing significant savings on Founders' annual support fees. We believe that robust support of this new electronic archiving service – very early on – will send an important signal to all constituents in the scholarly community that the long-term preservation of born-digital content is important, and is being dealt with seriously by those most affected. (Details of the publisher and library contributions are available from [www.portico.org](http://www.portico.org).)

### **Current Status and Looking Forward**

As of March 10, 2006, seven publishers have committed more than 3,000 journals to the Portico archive. Participating publishers include Elsevier, John Wiley & Sons, Oxford University Press, American Mathematical Society, UK Serials Group, Berkeley Electronic Press, and Symposium Journals (UK). Although Annual Archive Support payment levels for libraries were only recently announced at the JSTOR Participants Meeting at the ALA Midwinter conference in San Antonio, we are very encouraged by the library (and publisher) response to date. Some libraries have already signed an archive license agreement and more than two dozen have given us a verbal commitment to support the archive. We are in discussions with a large number of publishers – commercial, university presses, and scholarly societies – and are encouraged by how these are progressing. As publisher and library participation in Portico grows, we will keep the community informed via updates to the Portico website ([www.portico.org](http://www.portico.org)).

We will also share with the community lessons we gain as we continue to learn more about the electronic archiving challenge. Our collective understanding of the nature of this challenge will continue to grow and evolve over time, and Portico will need to evolve and change as well. We will rely upon input from the community to guide us through these changes, just as we have relied upon community engagement to shape

Portico initially. The Portico Advisory Committee, which is comprised of leaders from across the spectrum of organizations involved in scholarly communication, will continue to be an important part of this engagement, but we are pleased to also be involved in the Library of Congress' National Digital Information Infrastructure Preservation Program.<sup>ix</sup> We look forward to continuing to work with these types of active efforts, even as we focus directly on our primary responsibility to preserve the scholarly literature being entrusted to us.

---

<sup>i</sup> Average 2003-04 library expenditures for electronic materials was \$2,718,015. See Mark Young and Martha Kyrillidou, *ARL Statistics 2003-04* (Washington: Association of Research Libraries, 2005).

<sup>ii</sup> [www.arl.org/arl/pr/presvejrnl0ct05.html](http://www.arl.org/arl/pr/presvejrnl0ct05.html) Accessed March 13, 2006.

<sup>iii</sup> Publishers participating in the Portico pilot included: the American Economic Association, American Mathematical Society, American Political Science Association, Association of Computing Machinery, Blackwell, Ecological Society of America, National Academy of Sciences (PNAS), Royal Society, University of Chicago Press, and John Wiley & Sons.

<sup>iv</sup> <http://www.rlg.org/legacy/longterm/repositories.pdf> Accessed February 28, 2006.

<sup>v</sup> Migration, which involves transitioning content from one file format to another as technology evolves and file formats become obsolete. Two other common preservation strategies include emulation, which involves strategies to make future technology mimic technology of earlier generations, and byte preservation, which involves simple storage of an unmodified stream of digital bytes without special provision for future display or functionality.

<sup>vi</sup> Details of the Journal Archiving and Interchange Document Type Definition created by the National Center for Biotechnology Information of the National Library of Medicine are available at <http://dtd.nlm.nih.gov/> Accessed March 13, 2006.

<sup>vii</sup> Additional details regarding Portico's archival approach are available from Evan Owens, "[A Format-Registry-Based Automated Workflow for the Ingest and Preservation of Electronic Journals](#)," November 8, 2005, Digital Library Federation Fall Forum, Charlottesville, VA. (<http://www.portico.org/about/community.html>). Accessed March 13, 2006.

<sup>viii</sup> [http://www.arl.org/osc/EjournalPreservation\\_Final.pdf](http://www.arl.org/osc/EjournalPreservation_Final.pdf) Accessed March 13, 2006.

<sup>ix</sup> The members of the Portico Advisory Committee are: John Ewing, American Mathematical Society; Kevin Guthrie, Ithaka; Daniel Greenstein, University of California; Anne R. Kenney, Cornell University; Clifford Lynch, Coalition for Networked Information; Carol Mandel, New York University; David M. Pilachowski, Williams College; Rebecca Simon, University of California Press; Michael Spinella, JSTOR; Suzanne Thorin, Syracuse University; Mary Waltham, [www.MaryWaltham.com](http://www.MaryWaltham.com); Craig Van Dyck, John Wiley & Sons, Inc.