Scaling Up and Scaling Out: Leveraging Preservation Infrastructure and Experience to Benefit the Community

Kate Wittenberg*, Amy Kirchhoff**, and Sheila Morrissey***

*Managing Director, Portico, 151 E 61st St. New York, NY 10065,
Kate.Wittenberg@portico.org
http://www.portico.org
**Archive Service Product Manager, Portico, 100 Campus Drive, Suite 100, Princeton NJ 08540, Amy.Kirchhoff@portico.org
http://www.portico.org
***Senior Research Developer Portico, 100 Campus Drive, Suite 100, Princeton NJ 08540,
Sheila.Morrissey@portico.org
http://www.portico.org

Abstract. Now well into its eleventh year as a preservation archive and service, Portico's preservation infrastructure – hardware, software, and key data and metadata models and definitions – has been subject to a continual process of review and revision that makes it possible for us to leverage our work to benefit the broader community. Since 2012, Portico has been delivering e-journal content to the British Library as part of their legally mandated deposit program, enabling the British Library to benefit from existing preservation expertise to manage the normalization of e-journal content. In this paper we will discuss Portico's development, the scaling up of our preservation capacity, the challenges and opportunities in our partnership with the British Library, and the value of leveraging and sharing existing preservation infrastructure, skills, and experience for the good of the broader community.

1 Introduction

Portico is a digital preservation service provided by ITHAKA[1], a not-for-profit organization with a mission to help the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Created in 2002, Portico was founded to build a sustainable digital archive to serve the academic community to enable publishers and libraries to be secure in the long-term preservation and accessibility of the e-journals being licensed, and to realize tangible benefits as they transitioned to greater reliance on digital content. Portico began as a project funded by The Andrew W. Mellon Foundation to further its seminal E-Journal Archiving Program. Today, with over 25 million articles preserved, Portico is among the leading digital preservation services in the world.

[1] http://ithaka.org/

Our approach to digital preservation is comprehensive—combining long-term content management and organizational commitment with a philosophical dedication to addressing the needs of tomorrow's scholars. Portico preserves content through a format-based archive management strategy. The key points of this strategy are: identifying significant preservation metadata at the initial point of preservation, and conservative and pragmatic migration of content at the point where such an activity is both safe and necessary. While the archive is dark, publishers and libraries are provided with audit privileges that allow them to review the status of content. Content in the archive becomes "light" for faculty, staff, and students at participating libraries whenever a trigger event occurs (cessation of a publisher's operations; discontinuation of a title by a publisher; back issues no longer offered by a publisher; or catastrophic and sustained failure of a publisher's delivery platform). In addition, the majority of the titles in Portico are available for post-cancellation access if needed. Upon receipt of a claim from a participating institution and confirmation of the past subscription status by the publisher, campus-wide access is provided to the requesting participating library.

For Portico, the key goals of digital preservation include:

- Usability—the intellectual content of the item must remain usable via the delivery mechanism of current technology
- Authenticity—the provenance of the content must be proven and the content an authentic replica of the original
- Discoverability—the content must have logical bibliographic metadata so that it can be discovered by end users
- Accessibility—the content must be available for use by the appropriate community

To meet these goals, we have defined and follow exacting standards and processes for content management, maintenance, and replication of the archive; we conduct self-checks and third-party archive certifications to ensure quality and security; and we maintain a delivery system and services to provide access to users in a manner that is easy to use and integrated with other online resources.

Three years ago, we began a substantial transformation of Portico's internal environment with the goal of building a premier content management organization with robust and flexible platforms and processes, with increased efficiencies and economies of scale. Part of this development involved expanding the archive's capability to handle new content types. Portico extended its content model to accommodate e-books and digitized materials (newspapers, books, documents, pictures, etc.), modified its preservation metadata schema, and migrated roughly 15 million METS[2]-based metadata files to a new preservation metadata format. In addition to these internal infrastructure and technology developments, Portico underwent a successful audit that made it the first CRL certified Trustworthy Digital Repository. Portico also submitted the winning proposal to become the legal deposit service for the British Library,

[2] http://www.loc.gov/standards/mets

initiating the development of a new type of preservation service, and demonstrating the viability of Portico preserved content outside of the Portico system. As a result of this work, Portico is positioned as a trusted and sustainable service in the rapidly changing scholarly communications environment, and is able to use its infrastructure and experience to benefit the broader preservation community.

2 The Current Environment

We recognize that Portico works within a rapidly-changing environment. and that we must be aware of these changes in order to respond effectively. Scholarly communication, particularly book publishing, continues to evolve in the digital space, along with an associated need for new approaches to preservation of the material (in both old and new content types) coming from traditional scholarly publishers, but also from libraries and individual scholars involved in the creation of digital research. In addition, there has been a steady rise in the pressure on scholarly publishers for open access to scholarly research, and a need for innovative and effective approaches to the sustainability, persistence, and preservation of this content.

Research libraries are also undergoing important changes. They have for some time been pressured by the combined challenges of shrinking resources resulting from the stressed macro economy, and by the disintermediation of the role of the librarian caused by new digital modes of scholarly publishing and research. The scholarly literature increasingly acknowledges a professional change in focus to “connections, not collections.” Libraries are, jointly and individually, developing strategic plans whereby they take the lead in providing infrastructure, tools, and support for faculty research workflows. Libraries have taken the initiative in implementing several projects focused on large-scale, substantive collaboration. Hathi Trust is perhaps the most fully developed exemplar of this trend toward resource pooling replacing local work. They have made great strides toward an infrastructure that includes digitization, aggregation, and preservation of non-digital content, acquisition, access, and preservation of e-journals, and provision of a university press digital publishing platform.

Another change is the increasing engagement by faculty in digital initiatives in the scholarly domain, including challenges to authoritative channels of scholarly communication, collaboration, research, and instruction. Digital humanities centers have developed rich corpora of digitized and born-digital literary and historical source documents, both primary and secondary, often enriched with mark-up and hosted within a framework of tools for discovery, for visualization, for both “close” and “distant” (machine assisted) reading, and for crowd-sourced enrichment of the collection. These initiatives raise critical questions concerning the long-term accessibility of non-traditional scholarship such as computer models, data, and the human interaction with these materials and software. Increasing numbers of scholars are trying to alter the “sociology” of scholarly communication, collaboration, and evaluation (including the factors affecting tenure decisions), and are looking to associate scholarly distinction not with publication, but instead with impact, including reception via social media. Faculty and researchers are actively seeking collaboration with the research library community to establish permanent homes for access to, and preservation of, these collections.

As we think about our role as a preservation organization, we see many opportunities to engage usefully and productively with these developments. In order to do so, however, we will need to think innovatively about our internal systems as well as our external relationships with the community as a whole, and with outside initiatives that intersect with our work. We must now see the possibility of productive partnerships and collaborations that will provide critical services to an increasingly complex community of content.

3 Scaling Up Infrastructure: Technical Challenges of Preservation at Scale

As noted in the call for papers for this conference[3], the digital universe in India alone is expected to increase by more than an order of magnitude by 2020. The capacity and capabilities of the digital preservation community will have to keep pace with this accelerating demand. As has been the case for many of our institutional colleagues in digital preservation, we at Portico have found that this reality of ever-increasing scale means a continuing process of reflection and re-invention, both of our technical and of our organizational infrastructure.

What are the technical challenges of such dynamic growth? As Clay Shirkey said, in reporting the results of the United States Library of Congress-sponsored Archive and Ingest Handling Test,

> Scale is a mysterious phenomenon -- processes that work fine at one scale can fail at 10 times that size, and processes that successfully handle a 10-times scale can fail at 100 times. […] Institutions offering tools and systems for digital preservation should be careful to explain the scale(s) at which their systems have been tested, and institutions implementing such systems should ideally test them at scales far above their intended daily operation, probably using dummy data, in order to have a sense of when scaling issues are likely to appear. [7]

Portico is now well into its eleventh year as a preservation archive and service. As part of its institutional mission and definition, Portico's preservation infrastructure – hardware, software, and key data and metadata models and definitions – has been

[3] http://www.ndpp.in/APA-DPDTR-2014/

subject to a continual process of review and revision, consonant with a disciplined understanding of normative software system lifecycles, and of the ever-changing information environment in which we all find ourselves. In the course of that process, Portico has undertaken two sorts of major system migrations centered on issues of scale: a refactoring of the workflow system ("ConPrep") that ingests publisher content streams to scale it up by a factor of 60 to a current capacity of approximately one million files per day, [5] and the migration of over 15 million preservation metadata records to a new, semantically richer, syntactically leaner metadata model. [4]

In scaling up the capacity of Portico's workflow system, a good deal of analysis went into determining what optimizations, or refactoring, of the original system design were necessary to enable us to take advantage of hardware scaling with newer, faster, and less expensive systems and storage devices. A number of factors placed pressure on the original design [5] [1], including:

- A performance bottleneck in the system component responsible for persisting workflow information to the database
- The total size of the content, which caused performance degradation of disk drives as the size of disk volumes grew, and consequently the length of time it took to perform fixity checks on archive content
- The sheer number of files that make up individual archival components (a journal article, an e-book), which resulted in significant overhead in per-file reading and writing, and in the overhead of loading files into the content management system
- The lack of standard submission packaging for electronic journal and other content, resulting in the need for batches with different characteristics (for example, batches comprising a single journal issue with a large number of files, and other batches with multiple issues, each with smaller numbers of files), with different processing requirements, making varying capacity demands, and necessitating the intelligent automated management and balancing of batch and process scheduling

Portico learned many lessons about what it understands to be the ongoing challenge of system refinement and refactoring to keep pace with increases in scale during the process of system migration in 2007. We learned about the need to gather data, to analyze and experiment, and rigorously to test our changes – also at scale – both before and throughout the migration process. We learned to look for the consequence of optimization of one part of the system on other parts, and to balance all parts (hardware, software, people and processes) of the system as a whole. Increasing the capacity of one part of the system required that we create tools to increase the scope and capacity of automation of such system administration tools as cleanup and maintenance, logging and log management, and, perhaps most importantly, in the automation of quality assurance of the content being manipulated, ingested, and managed by the archive.

These were all lessons that stood us in good stead a few years later, when, in anticipation of new content types (electronic books, digitized collections) coming into the archive, and aware of refinements to, and maturing of, standards and practices in the digital preservation community (for example, the publication of the PREMIS data dictionary), Portico undertook a revision of its content model. This in turn resulted in new schema for Portico's preservation metadata, and the consequent migration of over 15 million already-existing metadata files to the new content model. As we noted in describing this migration, preservation metadata plays a key role in the archive:

> The archive's preservation activities are made manifest through the preservation metadata generated and collected throughout the life cycle of a preserved object. In Portico's case, these data can be generated during processing in ConPrep, at ingest to the archive, and as preservation activities take place thereafter.
>
> This meant that nearly every part of the system was likely to be "touched" in some way by the metadata migration. It meant as well, as indeed Portico's experience in scaling up ConPrep had demonstrated, that the migration would need to be carefully thought through, documented, managed, and coordinated amongst staff who would also be engaged in other work. [4]

Again, we undertook a careful process of analysis and planning, testing and tuning, to accomplish the migration.

4 Scaling Up Infrastructure: Organizational Challenges of Preservation at Scale

The Portico experience of scaling up its system capacity in particular gave us a view of the human impact of scale. Particularly affected was the Portico production staff, who were responsible for superintending the loading of content, for performing quality assurance on new content streams, and who performed problem resolution during processing. [5] Portico was challenged to accomplish an order of magnitude increase of system capacity, without necessitating an order of magnitude increase in the number of production staff required to handle this scaled-up capacity. Accomplishing this meant developing more automated tools, redesigning the graphical user interface, and refining and redefining staff roles and responsibilities.

The organizational challenge of preservation at scale is something that Portico shares with many others working in digital preservation. Realizing the growing awareness of, and experience with, the complexities of scale, Portico, along with colleagues from the National Library of the Netherlands, organized a workshop on preservation at scale to be held as part of IPRES 2013. Over thirty people from seventeen countries attended this workshop, including representatives from national libraries of Sweden, the Netherlands, France, the United Kingdom, China, and Germany, as well as from many university libraries, NGOs, and digital preservation consortia. As indicated in the report on the workshop [3], there were a number of broad categories addressed by the workshop speakers and participants, including

- the technological adaptations to collect and preserve ever-increasing amounts and varieties of digital content while taking advantage of new advances in both hardware and software
- the institutional adaptations, sometimes including new institutional self-definition, necessitated by the increasingly large role that the preservation of digital, as opposed to analog, objects has come to play in memory institutions
- quality assurance at scale and across scale, often at cross-purposes with the large-scale automation required to handle increasingly large amounts of digital content
- the scale of the long tail of unpreserved digital content,
- economies and the sometimes unexpected diseconomies of scale

It was particularly interesting to hear reports on the institutional challenges and responses to those challenges from the National Library of the Netherlands [8], the British Library [6], and Harvard University Library [2].   For these institutions, the sheer scale of the acquisition, management, and preservation of digital content has moved digital preservation from what might be called a satellite activity, to the core of each library's mission and self-definition. This in turn has often resulted in dramatic organizational restructuring. As the workshop report notes,

> This shift -- from relegating the preservation of digital content to an organizational sub-unit, to ensuring that digital preservation is an organization-wide endeavor -- is challenging, as it requires changing the mindsets of many in each organization. It has meant making choices and reallocation of resources from other activities, recognizing that the organization cannot do everything. It has necessitated strategic planning and budgeting for long-term sustainability of digital assets, including digital preservation tools and frameworks – a fundamental shift from one-time, project-based funding. It has meant comprehensive review of organizational structures and procedures, and has entailed equally comprehensive training and development of new skill sets for new functions. [3]

5 Scaling Out: Collaborations and Centers of Excellence—The British Library Partnership

As part of our mission, Portico explores new opportunities that address the emerging preservation needs of the broader community. These opportunities may include forming partnerships with key players and stakeholders to advance an overall, community preservation agenda and agreed-upon goals. As one step in this area, Portico is engaged in research related to the emerging preservation needs of researchers, including the preservation of data that supplements and supports journal articles and monographs, and the preservation of the links between published content and data.

We recognize that as an established, trusted, and sustainable preservation partner for publishers, libraries, and individual scholars, Portico can and should play a key role in leveraging our work in order to support the print-to-digital transition for the global community of libraries and individual scholars. We need to recognize opportunities and initiate projects that address current and emerging preservation needs of the community as a whole.

As part of this work, we are involved in a partnership to provide preservation services for the British Library's legal deposit program. In 2012, Portico and the British Library jointly worked on a pilot project to identify the tools needed to automate the transfer of content from Portico to the British Library and the communication channels needed to identify and correct problems in the content transfer. In early 2013, the U.K. Parliament approved legal deposit legislation requiring the deposit of all content electronically published in the U. K. with the national deposit libraries, and in April 2013, Portico began to deliver preservation files for e-journal content to the British library as part of this ongoing deposit program. Through 2013, Portico will be delivering content for 3 publishers and nearly 1500 journals (Portico has delivered over 108,000 articles to the British Library between April 2013 and November 2013).

The relationship is highly collaborative and allows the British Library to leverage the existing Portico technical infrastructure and staff expertise to manage the normalization of e-journal content to meet their legal deposit obligations. In addition, the British Library may also leverage the existing Portico publisher relations infrastructure and staff expertise to manage the content conversations with the publishers. The publishers that choose to participate in the program make arrangements with the British Library and sign an addendum to their existing agreement with Portico or, if they are not a Portico publisher, sign an agreement specific to the British Library arrangement with Portico.

The British Library project has influenced some changes in the Portico processes. Due to the long-term nature of Portico's preservation model, there has historically not been an urgent need to process and ingest content into the archive immediately upon its publication. However, the British Library needs to ingest the content into its preservation system quickly upon publication in order to make that content available in the Library reading rooms and to deliver the content to the other deposit libraries in the U.K. For the British Library publishers, therefore, Portico has increased the speed and flexibility of its receipt and ingest process. As with any electronic system, managing the limited set of special cases (issues published without an issue number, issues published and never delivered to Portico, etc.) requires more time from Portico staff than the large amount of content that simply processes straight through. As Portico is providing this service simultaneously for the content it is preserving and the content it is delivering to the British Library, the overall costs to the community are reduced.

Portico and the British Library have collaboratively designed a delivery system that relies on a number of standards, including:

- BagIt[4] for packaging and transmission of content
- Journal Article Tag Suite (JATS)[5] for e-journal XML mark-up
- Dublin Core[6] for metadata mark-up

We have learned a great deal about the challenges and the benefits of this arrangement, which we can now share with the broader community. Our collaborative work with the British Library has involved defining and architecting the Portico systems and exploring various content transformation and delivery options, and then making final choices together. We have addressed challenges around coordinating project management, working with different software development styles, and communication between the geographically distant teams. We believe that the lessons we have learned through this project will be useful to the community as many consider how to manage new preservation needs while realizing economies of scale.

First, it is important that we have identified ways to ensure that the library community is not paying to develop the same tools multiple times. The Portico/British Library partnership allows each party to focus on separate activities and leverages work each has already done.

Second, we have learned that communication, rather than technology, is the most common challenge for this sort of collaborative project. Regular communication and close cooperation between the Portico and British Library team members on a geographically and organizationally complex project has been a key factor for success. In addition, the teams had to agree on secure and efficient protocols for moving content back and forth between the two organizations. With a clear workflow process in hand, both organizations could independently develop their work in parallel.

Finally, we have come to understand that by leveraging and sharing Portico's existing infrastructure, experience, and skills, it is possible for us to create preservation solutions that provide significant value to others in our community.

References

1. Cheruku, Vinay: Preservation at Scale Issues and Solutions: A Portico Experience. Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at http://ipres2013.ist.utl.pt/workshop5_presentations/Vinay-Pres%20at%20Scale.pdf
2. Goethals, Andrea: Challenges and Lessons Learned Migrating an Entire Repository. Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at

[4] http://sourceforge.net/projects/loc-xferutils/
[5] http://jats.nlm.nih.gov/
[6] http://dublincore.org/documents/dces/

http://ipres2013.ist.utl.pt/workshop5_presentations/DRS2_Challenges_and_Lessons_Learned.pdf

3. Kirchhoff, Amy, Morrissey, Sheila, Ras, Marcel: Workshop Report: Preservation at Scale, Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at http://ipres2013.ist.utl.pt/workshop5_presentations/PreservationAtScaleWorkshopReport.pdf
4. Morrissey, Sheila, Cheruku, Vinay, Meyer, John, Stoeffler, Matthew, Howard, William, Kadirvel, Suresh : Migration at Scale: A Case Study. Presented at iPRES 2012, Toronto, October 1-5, 2012. In IPRES 2012 Proceedings of the 9th International Conference on the Preservation of Digital Objects, Reagan Moore, Kevin Ashley, and Seamus Ross (eds.), pp. 97-105. Available at https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf
5. Owens, Evan , Cheruku, Vinay, Meyer, John, Morrissey, Sheila : Digital Content Management at Scale: A Case Study from Portico. Presented at DLF Spring Forum, Minneapolis, April 28-30, 2008. Available at http://old.diglib.org/forums/spring2008/presentations/Owens.pdf
6. Pennock, Maureen: Organising Preservation at Scale: The British Library's Digital Preservation Strategy. Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at http://ipres2013.ist.utl.pt/workshop5_presentations/iPRES2013_pennock1-r.pdf
7. Shirkey, Clay: Library of Congress Archive Ingest and Handling Test (AIHT) Final Report. June 2005, page 26. Available at http://www.digitalpreservation.gov/documents/ndiipp_aiht_final_report.pdf
8. van Wijk, Caroline, Ras, Marcel: Finding the Balance: Technique, organisation, roles, responsibilities and flexibility. Presented at The Preservation at Scale Workshop at iPRES 2013, Lisbon, September 2-6, 2013. Available at http://ipres2013.ist.utl.pt/workshop5_presentations/Finding%20the%20balance-iPRES2013.pdf