BECOMING A CERTIFIED TRUSTWORTHY DIGITAL REPOSITORY: THE PORTICO EXPERIENCE

Amy Kirchhoff **Eileen Fenton** **Stephanie Orphan** **Sheila Morrissey**

Portico
100 Campus Drive, Suite 100
Princeton, NJ 08540

ABSTRACT

The scholarly community's dependence on electronic resources is rapidly increasing and those electronic resources are increasingly preserved in digital repositories or other preservation services. Whether locally hosted at libraries, collaboratively hosted between institutions, or externally hosted by a third party, one method for these digital repositories to take to assure themselves and their communities of their soundness is to be audited and certified by impartial organizations. Such independent organizations with staff experienced in executing audits and certifications can represent the interests of the academic community. Such staff will have the time and skills required to perform a thorough review of the methodologies and policies of each digital repository.

Over the course of 2009, the Center for Research Libraries (CRL) audited Portico, a third party preservation service. At the conclusion of the audit, CRL certified Portico as a trustworthy digital repository. The audit was a lengthy, productive experience for Portico. We share the experience here both to impart the depth of the audit and to inform other organizations of what steps might be involved should they choose to be audited and certified.

1. INTRODUCTION

Over the course of 2009, the Center for Research Libraries (CRL) audited the Portico preservation service. The audit formally concluded in January 2010, when CRL certified Portico as a trustworthy digital repository. Portico is the first preservation service so certified by CRL.

> *"The Center for Research Libraries (CRL) conducted a preservation audit of Portico (www.portico.org) between April and October 2009 and, based on that audit, has certified Portico as a trustworthy digital repository. CRL found that Portico's services and operations basically conform to the requirements for a trusted digital repository. The CRL Certification Advisory Panel has concluded that the practices and services described in Portico's public communications and published documentation are generally sound and appropriate to both the content being archived and the needs of the CRL community. Moreover the CRL Certification Advisory Panel expects that in the future, Portico will continue to be able to deliver content that is understandable and usable by its designated user community."* [1]

Portico (www.portico.org) is a not-for-profit digital preservation service providing a permanent archive of electronic journals, books, and other scholarly content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. In May 2010, there were nearly 15 million articles and 2,000 e-books preserved in the Portico archive with over 10,000 journals, 30,000 books, and 10 collections of digitized historical content committed to the archive. We anticipate that an additional 1.5 to 2 million articles, tens of thousands of e-books, and several d-collections (digitized historical collections, such as historical newspapers) will be preserved in the archive every year.

CRL (www.crl.edu) is an international consortium of university, college, and independent research libraries.

The CRL audit of Portico extended through ten months from April 2009 to January 2010. It was the first preservation audit Portico has undergone. Ultimately, this audit was a collaborative and productive learning experience.

As an element of certification, CRL assigned Portico levels of certification in three categories: organizational infrastructure; digital object management; and technologies, technical infrastructure, and security. Portico's score for each category is given below in Table 1. (The numeric rating is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level.)

Category	Portico Score
Organizational Infrastructure	3
Digital Object Management	4
Technologies, Technical Infrastructure, Security	4

Table 1. Portico Certification Scores

In addition to the formal scoring, Portico and CRL agreed that over time Portico would address some of the

concerns CRL highlighted in the written audit report[1] and in informal discussions with Portico (for example, improving the Portico roles and responsibilities documentation).

Portico benefitted from the audit in practical and tangible ways. Our preparation for the audit, which included collecting and updating documentation, made it easy to provide this documentation to other parties subsequent to the audit. The most significant benefit is the assurance regarding the viability, the integrity, and the effectiveness of our preservation approach that only such a comprehensive, objective, third-party review can provide.

2. REASONS TO BE AUDITED

An audit is "an evaluation of a person, organization, system, process, enterprise, project or product" [5] and certification is "the confirmation of certain characteristics of an object, person, or organization … this confirmation is often … provided by some form of external review, education, or assessment." [6] The CRL preservation audit and subsequent certification of Portico as a trustworthy digital repository was just that, an external review and evaluation of Portico.

The yearly statistics produced by the Association of Research Libraries (ARL) show that every year the scholarly community becomes more dependent on electronic content (see Figure 1). Indeed, by 2008 the ARL institutions were spending over 50% of their library materials expenditures on electronic resources – resources that by their very electronic nature are not preserved on the shelves of the library itself.

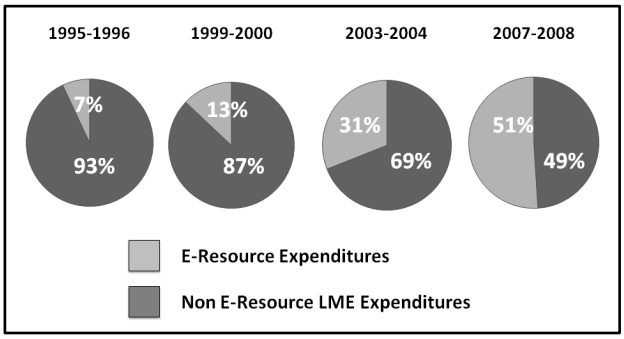

Figure 1. ARL E-Resources Expenditures[2]

Portico preserves an ever growing portion of these digital resources the scholarly community relies upon, and as such we felt it was imperative that we undergo a formal third party assessment and certification process to assure ourselves, the ITHAKA board, and, most importantly, the scholarly community, that our preservation methodology, processes, and archive will secure the long-term preservation of the content in our care.

[1] http://www.crl.edu/sites/default/files/attachments/pages/CRL%20Report%20on%20Portico%20Audit%202010.pdf

[2] This chart is created from the publicly available figures in the ARL annual statistics at http://www.arl.org/stats/annualsurveys/arlstats/index.shtml

3. AUDIT METHODOLOGY

CRL based its audit process on the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), as well as other inputs of interest to the CRL community. These inputs included *"metrics developed by CRL on the basis of its analyses of digital repositories. CRL conducted its audit with reference to generally accepted best practices in the management of digital systems; the interests of its community of research libraries; and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada. The purpose of the audit was to obtain reasonable assurance that Portico provides, and is likely to continue to provide, services adequate to those needs without material flaws or defects and as described in Portico's public disclosures."* [1]

TRAC is a standard that was developed by experts within the digital preservation community. Its goal is to identify the criteria that define a trustworthy digital repository. It is important to be aware that TRAC and other digital repository audit methodologies, such as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), are designed to evaluate a repository against its own claims, not against a single standard set of measurements. *"At its most basic level an audit should assess whether a repository can meet its stated commitments–is it doing what it says it is doing?–and the criteria have to be seen within the contexts of the special archiving tasks of the repository."* [2] With such a focus on the context of the specific repository being evaluated against TRAC, two repositories with very different levels of documentation, and indeed with very different kinds of preservation goals, service level models, and guarantees, could both be certified, if the level of documentation at each repository supports that repository's individual purpose and public statements.

The CRL audit team consisted of two full-time CRL staff members and one CRL technical consultant. Guidance and advice on areas of concentration for all CRL digital repository preservation audits is provided to the CRL audit team by the CRL certification advisory panel, which represents the CRL membership and "*its community of research libraries and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada."* [1] The CRL certification advisory panel includes leaders in collection development, preservation, and information technology.

At Portico we made several important decisions early on in the audit process: 1) we agreed it was important to ensure that the CRL audit team understood our preservation philosophy, policies, and workflow, and 2) we would establish a primary contact for CRL throughout the process. The Portico archive service product manager, Amy Kirchhoff, coordinated the internal process and communicated externally with the CRL audit team, while many staff members of Portico and ITHAKA were involved in the audit process. In particular, the Portico senior research developer, Sheila

Morrissey, and publisher content coordinator, Stephanie Orphan, were heavily involved in audit preparations. CRL and Portico collaborated on the development of the timeline and logistics for the audit process. Over the course of several conversations, we worked together to identify what documents would be required.

Portico gathered documentation and expertise from all parts of the organization and provided CRL with five subject based portfolios of documentation. To aid this portfolio creation, we developed an internal document cross-referencing nearly all of Portico's documentation to the TRAC criteria. Shortly after receiving the documentation from Portico, the CRL audit team visited the Portico New Jersey office to witness and audit the steps Portico takes in its preservation process. Following the site visit, there was an ongoing dialogue between Portico and the CRL team as we worked to address their questions about our preservation process, policies and documentation. While the audit itself was quite rigorous, it was a productive and collaborative process.

3.1. Documentation

As with virtually all kinds of audits, the CRL digital repository assessment requires the repository to provide evidence to demonstrate how it meets the audit criteria. This evidence-based methodology is intrinsic to TRAC, *"in particular, appropriate documentation of all steps permits auditors to evaluate the digital long-term repository as a whole"* [2] and DRAMBORA, *"a range of evidence expectations are described within the audit tool, reflecting a belief that organizations must be able to demonstrate their ability to effectively manage their risks."* [3]

In support of this evidence-based methodology, we spent several months identifying documentation we had already written and cross-referencing it to TRAC. Before the site visit, Portico provided the CRL audit team with 1,225 pages of documentation organized into five portfolios:

- **Organization**: including items such as organizational charts, meeting notes, financial statements, documentation of surveys, and sample email conversations with participants
- **Policy**: including all Portico preservation policies
- **System Architecture and Content Model**: including several introductory presentations, and content model & information architecture documentation
- **Operations and Systems Development & Maintenance**: including content manifests, illustrative documents from Portico trigger events and instances of post-cancellation access, sampling of minutes from the weekly technology & operations meetings, documentation for major systems changes, Portico disaster recovery plan, documentation of the results of retrievals from backup, support contracts with external vendors, receipts for payment of cloud storage service fees, and documentation about fixity verification processes, including recovery in case of errors found on disk
- **Archive Interfaces**: including user and business requirements for the audit and access interfaces to the Portico preserved content and documentation about planned enhancements to the auditor interface

For these portfolios, Portico staff collected previously written documentation and reproduced that documentation in image form. In order to provide context to each document, Portico wrote introductions to precede most documents. We completed significant writing for the audit in the area of policies–many of Portico's policies were encoded in training classes and operational procedures (which were also provided to the CRL audit team). Preparing for the audit created an opportunity for us to consolidate our understood "policies" into formal policy documents.

After receiving the portfolios of Portico documentation, visiting Portico on-site, receiving sample articles exported from the Portico archive, and reviewing all of the information gathered throughout the audit process, the CRL audit team requested additional documentation from Portico, including:

- Samples of the "Portico Modification to Original Submission Information Packages or Portico Archival Units Form"–a document Portico uses for tracking purposes when it is necessary to modify content outside of the standard ingest workflow, for example if prior to ingest Portico will be replacing corrupted content with corrected content as provided by the publisher.
- Sample format action plans (format action plans are documents that describe how an organization will address the preservation needs of specific file formats) and turn over documents (which specify the format action plans for publisher-specific XML and SGML formats and publisher-specific packaging schemes).
- Lists of formats and file types accepted into the archive and any formats and file types not accepted. Portico accepts all file formats into the Archive and provided the CRL team with a list of all formats in the archive (files in the Portico archive are assigned a preservation level determined by the tools available to support the file format and the commitments made to the specific content (e.g. well-formed PDF files associated with e-journal articles are fully preserved, whereas ill-formed PDF files or executable applications are byte preserved)—as file format tool sets improve over time, the preservation levels assigned to specific files will be adjusted.)
- Brochures designed for library and publisher outreach, provided as PDFs.
- Example license agreements as exported from the archive.

- Relevant technical certifications earned by Portico.
- Documentation of any hardware and software changes. This information is encoded in the event records in the archival information packages preserved in the archive.
- Budgets and expense/revenue statements for 2005-2009.
- Sample communication to publishers regarding status of their content. Twice a year, Portico provides publisher participants with a report that includes general information about Portico status and specific information about that publisher's content in the archive.
- A sample publisher agreement annex in spreadsheet form. This document lists what content is committed to the Portico archive.
- An explanation of the process used to produce library-specific holdings comparison reports– these reports compare the holdings of the Portico archive to those of a specific library or portion of a library's collection.

3.2. Beyond Documentation

In addition to producing the documentation portfolios and providing additional documents on request, Portico engaged with CRL through numerous phone and e-mail conversations. While Portico and CRL had a number of conversations about audit logistics, the majority of the conversations were initiated by the CRL audit team as questions arose during their review of Portico-provided documentation and sample articles. Many of these questions required responses rich with information and we appreciated the opportunity to clarify Portico policies and practices.

We received general technical questions from the CRL audit team, including questions about the Portico information architecture, replication policies, and bit corruption tolerance. (Portico has a zero tolerance policy, which is not documented separately, but is reflected in the fixity verification documentation.) The CRL audit team reviewed the sample articles in depth, compared them to the content model documentation we provided, and developed a variety of article-specific questions about identifiers and other required (or not) descriptive metadata, article presentation for delivery, and content transformation.

The CRL team was quite interested in exploring and testing retrievals from the Portico archive. In order to address this concern, we explained that we frequently export content from the archive including regular exports to the archive replicas, the delivery site, and (at the time of the audit and in accordance with existing publisher agreements) to the Library of Congress. In addition, we perform a number of one-off exports to our participating publishers.

The CRL audit team was also particularly interested in the Portico holdings and ways for the community to gain detailed information about the specific contents of the Portico archive. We discussed tools such as: the audit web interface through which librarians and publishers may review archived content, the Portico holdings comparison tool that compares a library's holdings to the Portico archive, and the detailed Portico holdings lists.

CRL also had questions about the business and technical logistics of providing post-cancellation access, a service that Portico provides to participating publishers on an opt-in basis. The CRL team also inquired whether Portico receives DTDs and schemas from publishers and whether they are placed in the archive and we confirmed that these materials are received and preserved in the archive.

3.3. Audit Timeline

Portico was involved in audit preparation and the actual audit for approximately 16 months (from the fall of 2008 through January 2010), although the audit itself extended over 10 months.

Winter 2008-2009–During the early winter of 2008, Portico and CRL held initial discussions about the proposed timeline for the audit.

Spring 2009–We began the process of identifying and collating existing documentation from a variety of departments, including finance, human resources, legal, information technology, content management, user support, delivery, publisher relations, outreach, and operations. This documentation was distributed across many systems, including Talisma (a contact management system), SVN (a version control system), JIRA (a bug, issue, and project tracking system), the Portico intranet, a Wiki, shared drives, web servers, the public website, local drives, and email accounts. Portico also began work on a TRAC self-report documenting to what degree we met the 84 criterion in TRAC and describing the documents available to support our assessment. (This TRAC self-report is available on the Portico website in the Archive Certification area.) [1]

CRL announced the launch of the audit of Portico in March 2009 and in April, Portico submitted the TRAC self-report to CRL.

Summer 2009–Portico developed a policy document template and policy approval framework and began to document existing policies using the new template. We continued to create the five portfolios of documentation. In May, Portico and CRL agreed to the logistics of the audit and we learned who at CRL would be on the audit team and how the team would interact with the CRL certification advisory panel. We provided the CRL audit team with access to the Portico auditor website and received the schedule of documentation from CRL. In July, Portico and CRL finalized the agreement guiding the audit process. Portico also provided references for third parties that received data exports from Portico. Portico submitted the first portfolio of documentation,

[1] http://www.portico.org/digital-preservation/wp-content/uploads/2009/10/CRL-Audit-Portico.FINAL_.pdf

the organizational portfolio, to CRL on August 4th. The four additional portfolios were submitted on August 13th.

On August 19th, the CRL audit team visited the Portico office in Princeton, New Jersey. As the CRL team was particularly interested in observing staff perform their normal activities, we arranged for the team to "follow" the content as it moved from one Portico unit to another. We started the day by attending the daily meeting between the technology and operations groups. Next we took the CRL team to talk with the publisher content coordinator who kicks off the business and analysis processes that begin after a publisher has signed a preservation license agreement. Next the CRL team spoke with the staff that develops publisher-specific tools that transform content to archival formats. The CRL team then spoke with members of the Portico systems team and attended a release coordination meeting. To end the visit, the CRL team met the Portico ingest team, where they witnessed the process of ingesting content into the archive and resolving problems with the content during the transformation process.

After the site visit, Portico staff wrote software to export the 200 sample articles requested by CRL from the archive and build a navigable set of HTML pages that would allow the CRL team to review the entirety of the archival information package for each article, including the archival metadata file (the current Portico auditor interface does not provide access to the archival metadata file or the publisher's original SGML or XML files). Portico made this set of pages available to the CRL via an FTP site, where we also made available the Portico tool registry, file format registry, business data objects (a database that maps Portico publishers to their titles, used for collection management purposes), and a set of 20 submission information packages (a submission information package is content as provided to Portico by the publishers before any archival processing).

Fall 2009–Portico and CRL interacted extensively and Portico provided additional documentation as requested (including job descriptions and additional financial information). In October 2009, we coordinated a conference call between the CRL audit team and the Library of Congress to allow the CRL team to learn from the Library of Congress about their experiences developing an export process with Portico and managing the receipt of content exported from the Portico archive.

Winter 2009-2010–Portico received the draft report from the CRL audit team and offered comments. In January 2010, CRL released the final audit findings, initially sharing the results with CRL members and then to the broader community.

Spring 2010–Portico and CRL continue to have conversations about areas of particular interest to CRL or in response to questions raised by the CRL membership.

3.4. Audit Costs

Over the course of the 16 months during which Portico was engaged in the audit process, many staff participated, including staff from library outreach, publisher outreach, legal, finance, user services, operations, and development. The Portico Archive Service Product Manager invested the most time, approximately four months of work. Combined, other staff contributed another four months of work. This staff cost was funded out of Portico's operating budget. Ongoing communications with CRL and the regular updates will also be funded out of Portico's operating budget. Portico will integrate addressing the concerns raised during the course of the audit into day-to-day operations. We believe the regular updates that must occur every two years will require significantly less staff time than the initial process.

3.5. Ongoing Audit Activities

The CRL report on Portico audit findings outlines concerns the CRL had on 12 of the 84 TRAC criteria [1]. The CRL team provided Portico with additional comments by email and phone. The concerns range from documentation discrepancies (e.g., discrepancies found between Portico job descriptions and Portico policy documentation) to very specific requests for a software and hardware patch register to more general concerns about usability. Portico is developing a road map that will allow us to address these issues over time. We remain in contact with CRL on areas of mutual interest (for example, how to share holdings information).

As appropriate, Portico has already addressed some issues identified in the CRL report. The CRL report identified concerns with the opaqueness of the Portico holdings comparison results and we recently rewrote the Portico holdings comparison tool such that we now provide summary information in a more intuitive layout with each comparison. Also, the CRL report identified a concern that Portico is short of archiving a "critical mass" of journal content. Eileen Fenton, the Portico managing director, participated in a recent ALCTS meeting hosted by Martha Brogan, the Chair of the CRL Certification Advisory Panel, at the ALA 2010 Annual Conference. The purpose of this meeting was to discuss the corpus against which any measurement of critical mass should be made.

4. LESSONS, SURPRISES, AND BENEFITS

While the audit entailed a substantial amount of work for Portico, the interactions with the CRL audit team were pleasant, productive, and beneficial. The CRL audit team was extremely thorough and reviewed in great detail all documentation and samples we provided, the Portico website, and the audit and access interfaces. We appreciated their deep interest in learning about the Portico processes. One substantial benefit from this process is simply the opportunity for external review and validation of the approach and processes employed by Portico in pursuit of our preservation work.

Early in the process, Portico decided it was important to ensure that the CRL audit team understood our

preservation philosophy, policies, and workflow. This decision to emphasize education and deep understanding had a large impact on the amount of effort required to complete the audit. Rather than collect documentation and forward it to CRL piece meal, we identified existing and missing documentation, collected and wrote documentation, collated it into portfolios, and wrote cover notes to nearly each document. The logistics of this manual process were time consuming. In the end, the process served Portico well.

It is difficult to measure what impact the CRL certification of Portico has made on decisions others make in regard to Portico participation. Portico's certification has been a point of conversation within discussions we are in with the National Library of Medicine in regard to whether or not Portico may be considered an acceptable archive in regards to Medline indexing (currently, the only acceptable archive is PubMed Central). [1]

Portico has benefitted in many ways from going through the audit process. We frequently interact with members of the community and respond to requests for information. We have been able readily to share materials collected and documented during the audit process as part of these dialogues. Another benefit arose from the CRL audit team's interest in speaking with a Portico data export partner. As a result we held debriefing conversations with each of our data export partners. These conversations helped us better define the inter-organizational aspects of a data export and ways we can bring Portico's data transformation expertise to questions that might arise during our partners' work with preservation formats and packaging.

It would benefit managers of repositories of all sizes to evaluate their repository against TRAC or perform a self-assessment of risk via DRAMBORA. Whether any individual repository should be audited by a 3rd party, such as CRL, will depend upon the preservation commitments that repository has made, the uniqueness of the content it preserves, and the importance of the content to the repository's designated community (the community served by the repository [3]). Repositories at a smaller scale than Portico and with a more limited community or preservation commitment will, perforce, not have the same level of documentation as Portico. Whether a 3rd party preservation audit is required of any given repository is a decision that must be made by the community served by that repository.

The greatest benefit to Portico was simply the reassurance to Portico and the ITHAKA Board, to the publisher community, to the library community, and to the greater academic community, that the Portico archive was being rigorously examined by an external party. Portico provides auditor privileges to a maximum of four librarians from each participating library and to representatives at each participating publisher (librarian auditors may audit the entire archive and publisher auditors may audit their own content), but it is important to supplement this independent and individual audit activity with a more extensive and systematic approach, as demanded by a TRAC-oriented audit.

To ensure that Portico's certification remains current, Portico will provide CRL with updated documentation every two years and will continue dialogue with CRL on a variety of topics, including what content Portico should target as high priority for preservation. We are looking forward to an ongoing and active dialog with CRL.

[1] http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a2

5. REFERENCES

[1] The Center for Research Libraries, *"CRL Report on Portico Audit Findings"*, Chicago, Illinois, 2010, http://www.crl.edu/sites/default/files/attachments/pages/CRL%20Report%20on%20Portico%20Audit%202010.pdf (accessed May 4, 2010).

[2] The Center for Research Libraries & OCLC, *"Trustworthy Repositories Audit & Certification: Criteria and Checklist"*, Chicago, Illinois, 2007, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed May 4, 2010).

[3] Consultative Committee for Space Data Systems (CCSDS), "Reference Model for an Open Archival Information System (OAIS)", Washington, DC, 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed July 11, 2010)

[4] Digital Curation Centre & DigitalPreservationEurope, "*DCC and DPE Digital Repository Audit Method Based on Risk Assessment,* v1.0", http://www.repositoryaudit.eu/ (accessed May 4, 2010).

[5] Wikipedia contributors, "Audit," *Wikipedia, The Free Encyclopedia,* http://en.wikipedia.org/w/index.php?title=Audit&oldid=359674809 (accessed May 4, 2010).

[6] Wikipedia contributors, "Certification," *Wikipedia, The Free Encyclopedia,* http://en.wikipedia.org/w/index.php?title=Certification&oldid=358904216 (accessed May 4, 2010).