



Portico Final Report Draft

26 October 2006

Developed by Robin Dale, with input from Bernard Reilly.
Technical Analysis by, Sayeed Choudry, and Tim DiLauro

Note: The present report was produced as part of a test of the RLG/NARA Draft Audit Checklist for the Certification of Trustworthy Digital Repositories, and other metrics developed by the Center for Research Libraries under a grant from the Andrew W. Mellon Foundation. Because the metrics and methodologies applied are still in development, this report should be considered a definitive assessment of the repository described.

TABLE OF CONTENTS

1	HIGH LEVEL SUMMARY.....	1
2	EXECUTIVE SUMMARY	3
3	FULL REPORT	5
3.1	INTRODUCTION.....	5
3.1.1	<i>Repository type & background</i>	<i>5</i>
3.1.2	<i>Portico Philosophy</i>	<i>5</i>
3.1.3	<i>Organizational Structure (Brief Overview).....</i>	<i>6</i>
3.1.4	<i>Technical Architecture (Brief Overview).....</i>	<i>6</i>
3.2	OBJECTIVES, SCOPE, & METHODOLOGY.....	7
3.2.1	<i>Scope.....</i>	<i>7</i>
3.2.2	<i>Method of Work</i>	<i>7</i>
3.2.3	<i>Standards Against Which Audit Was Completed</i>	<i>7</i>
3.3	FINDINGS.....	8
3.3.1	<i>Organizational Analysis</i>	<i>8</i>
3.3.2	<i>Content Analysis</i>	<i>13</i>
3.3.3	<i>Technical Analysis</i>	<i>18</i>
3.3.4	<i>Vulnerabilities.....</i>	<i>30</i>
3.3.5	<i>Observations and Recommendations.....</i>	<i>31</i>
4	APPENDICES.....	32
4.1	COMPLETED PORTICO AUDIT CHECKLIST (22 PAGES)	
4.2	PORTICO RESPONSES TO ADVANCE TECHNICAL QUESTIONS	
4.3	PORTICO RESPONSES TO ADVANCE FINANCIAL QUESTIONS	
4.4	PORTICO FUNCTIONAL OVERVIEW	
4.5	PORTICO BOARD OF DIRECTORS	
4.6	ITHAKA BOARD OF TRUSTEES	
4.7	PORTICO JOURNAL ARCHIVE LICENSE AGREEMENT (LIBRARIES)	
4.8	PORTICO PUBLICATION LICENSE AGREEMENT (PUBLISHERS)	
4.9	PORTICO INITIAL FINDINGS POWERPOINT PRESENTATION	
4.10	PORTICO PARTICIPANTS MEETING AT ALA (JUNE 2006) HANDOUT	

1 High Level Summary

The audit of Portico, an electronic journals archiving service, took place 22-23 March, 2006. Among the goals of the test audit was to evaluate Portico to form an overall risk analysis as it relates to long-term access to scholarly electronic journals and other digital resources being managed by Portico. Using a draft of the RLG-NARA Checklist for the Certification of Trusted Digital Repositories as metrics or controls, as well as other documentation prepared exclusively for this project, Portico was evaluated on three aspects of the archive's operation:

- a. characteristics of the archiving service that affect performance, accountability, and business continuity;
- b. technologies and technical infrastructure employed by the archive; and
- c. archives processes and procedures adopted by the archive.

After the 2-day audit, findings of the audit revealed a relatively young enterprise with a new, developing archive in the very earliest stages of being populated with content. Because the Portico organization and system are still in development, it would be premature to render a judgment on its sustainability. Hence our findings and opinions on the archive must be considered qualified. The Portico system was just moving into production at the time of the audit, but indications are that the system is well-designed and managed, capable of providing the specific level of content functionality to which Portico is committed in its agreements with publishers and libraries. Portico's prospects to achieve its preservation goals look positive, although the service is a new model for the publishing and library world. There are certain sustainability risks related to the service's revenues and costs over the next five years. Portico will need to prove the worth of its service, and thus the viability of its business model, to the stakeholder communities in order to meet revenue projections and sustain itself.

Although some issues have been identified through the audit, many of these pertain to documentation and are indicative of the relative youth of the archiving service. These will need to be remedied in the near future to maintain transparency (a stated organizational goal) and enable change management as Portico continues to evolve. In the interim however, stakeholders in Portico should be encouraged by the capabilities of the archiving system demonstrated to date and in Portico's prospects for adequately managing electronic journal content for long-term use.

2 Executive Summary

In March 2005, the Andrew W. Mellon Foundation funded the Center for Research Libraries (CRL) Auditing & Certification of Digital Archives project, an endeavor to develop an audit and certification process for digital repositories and archives. Rigorous auditing and certification are necessary to determine the level of assurance that particular archiving arrangements provides to publishers/depositors and users, and to ensure that the valuable digital resources archives will continue to be available and functional over the long-term. As a part of the CRL project, a team of three auditors performed an audit and assessment of Portico, the new electronic archiving service. As a component of the project, Portico agreed to participate as a “subject archive,” one of three such archives to undergo a test audit as a part of the process development.

The audit of Portico took place 22-23 March, 2006. Among the goals of the test audit was to evaluate Portico to form an overall risk analysis as it relates to long-term access to scholarly electronic journals and other digital resources being managed by Portico. Utilizing a draft of the *RLG-NARA Checklist for the Certification of Trusted Digital Repositories* as metrics or controls, as well as other documentation prepared exclusively for this project, Portico was evaluated on three aspects of the archive’s operation:

- a. characteristics of the archiving service that affect performance, accountability, and business continuity;
- b. technologies and technical infrastructure employed by the archive; and
- c. archives processes and procedures adopted by the service.

After the 2-day audit, findings indicate a young enterprise with a new, developing archive in the very earliest stages of being populated with content. The Portico system was just moving into production at the time of the audit, making it premature for us to judge the likelihood of its eventual success. Hence our opinion of the archive must be considered qualified.

Nonetheless indications are that the system is well-designed and managed, and will eventually be capable of providing the functionality to which it is committed through Portico’s agreements with publishers and libraries. The outlook for Portico is generally promising, although it represents a new and untested model for the publishing and library world, being neither an entirely “dark” archive nor a regularly accessible collection. As a result there are certain sustainability risks related to Portico’s ability to meet revenue requirements and adequately control costs over the next five years. Portico will need to prove the worth of the service to the stakeholder communities in order to meet its revenue projections and sustain itself.

As expected, some vulnerabilities were identified in the course of the audit. Many of these pertain almost exclusively to high-level classes of technical issues including documentation and automation and are indicative of the relative youth of the service and archive. And while there is a tendency to postpone or minimize documentation in favor of developing the system for production, there is still a need to produce process and procedural documentation that lend to the transparency of the organization and service.

Documentation was identified as incomplete or lacking in the following major areas:

- Technical policies
- Procedures
- New requirements
- Replication plan
- Disaster plan and mitigation of flood or water threats to the archive; and
- Service continuity plan

The latter three issues are seen as critical for an organization such as Portico where a large community of stakeholders will potentially rely upon it not merely as an escrow or backup archive, but ultimately as a provider of access to the archived content. This should be remedied as soon as possible to engender publisher and subscribing institution confidence that Portico can meet what technological or environmental problems it encounters. Other, more minor issues revolve around issues of automation and use of available data. Portico's technical infrastructure is capable of recording and logging many aspects of the system, but there are currently few ways to exploit that information to consistency check the archive. Utilizing the logs in a more machine-enabled, routinized way would reduce the risk of data damage or loss and should be investigated as a development possibility.

It is also unclear, at this stage, how and to what extent Portico will be able to address the future costs of migration and service of the archive in the event of publisher failure or other "trigger events" that will move the organization into the role of providing access to the archived content. While Portico has in JSTOR a ready-made platform for delivery of electronic journal content, how the pricing model will change to support such service is not yet apparent, nor is it evident the extent to which that model will be acceptable to the existing user community.

These questions aside, however, it is clear that Portico has made a series of intelligent design and implementation decisions underpinning this new kind of archiving service. Portico has been built to utilize common hardware and software, but simultaneously created a specialized, centralized system capable of archiving e-journals in a way that is consistent with the desires of the community they support. Their preservation strategy of format normalization frontloads a great deal of the work and preservation "heavy lifting" that would otherwise need to take place far into the future. Choosing to tackle these issues at a time when they are closest to content creation – at point of deposit and ingest – is a wise decision for content such as electronic journals. Preserving the intellectual content versus the "look and feel" of originally published material is also a big step forward. Portico may be setting the preservation strategy for this kind of content for the foreseeable future.

At the time of the audit, very little content existed within Portico, although in the ensuing three months it moved into production and had ingested over 20,000 pages as of the end of June. Based on those rates of throughput, and assuming consistent, reliable ingest and the continued growth in interest among journal publishers, Portico expects to be well on its way to managing and providing access to thousands of titles and millions of pages by the end of calendar 2007. Barring unforeseen events between now and then, Portico should have assembled a considerable amount of critical content on which librarians, scholars and researchers will be able to rely.

In general, Portico seems likely to fulfill its mission of preserving scholarly journal literature published in electronic form and ensuring that these materials remain accessible to future generations of scholars, researchers, and students. Stakeholders in Portico will want to monitor its next few years closely. In that time period, Portico will need to shift revenue sources, grow its subscriber base, and scale up the system to continue ingesting the new content and backfiles already promised to them. The addition of content from new publishers will further contribute to this. While it could be a difficult period of growth for the organization, our audit suggests that Portico will soon be capable of providing subscribers reasonable assurance of access to significant journal literature for the near-term future. Stakeholders should be optimistic about the abilities of the archiving system to meet the demands precipitated by publisher failure or other such trigger event or perpetual access obligations.

3 Full report

3.1 Introduction

As a part of the Center for Research Libraries Auditing & Certification of Digital Archives project, a team of three auditors (Robin L. Dale, Tim DiLauro, and Marie Waltz) performed an audit and assessment of Portico, the electronic archiving service under development through Ithaka Harbors. The Andrew W. Mellon Foundation-funded project is engaged in developing a complete audit and certification process for digital repositories and archives. Rigorous auditing and certification are necessary to determine the level of assurance that particular archiving arrangements provide to publishers/depositors and users, and to ensure that the valuable digital resources archives will continue to be available and functional over the long-term. As a component of the project, Portico agreed to participate as a “subject archive,” one of three such archives to undergo a test audit as a part of the process development. The results of that test audit are the subject of this report.

3.1.1 Repository type & background

Portico began as the Electronic-Archiving Initiative launched by JSTOR in 2002 with a grant from The Andrew W. Mellon Foundation. The purpose of the Initiative was to build a model for a sustainable digital archive of scholarly literature published in electronic form. For more than three years project staff worked to develop the necessary archiving technology while engaging in extensive discussions with publishers and libraries about how to craft an approach that balanced the needs of publishers and libraries while generating sufficient funding to be self-sustaining. In 2004, the Electronic-Archiving Initiative became a part of Ithaka, a new non-profit organization with a mission to “accelerate the productive uses of information technologies for the benefit of higher education around the world.” In 2005, a new electronic archiving service resulting from the Initiative’s extensive collaborations with the community was launched under the name Portico and the direction of Eileen Fenton, Executive Director.

The Portico archive model builds from two core assumptions: first, material is preserved for eventual use and access and second, a completely and perpetually “dark” archive is not desirable. From these assumptions and drawing heavily upon extensive input from libraries and publishers, the Portico electronic archiving service has been shaped as follows: Portico will act as a centralized repository. Publishers will submit source files of scholarly electronic journals to Portico, and Portico will normalize these source files to an archival format and provide long-term archival management and format migration as needed. Libraries that support the archive will have campus-wide access to archived content when very specific conditions or “trigger events” occur.

3.1.2 Portico Philosophy

The mission of Portico is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future generations of scholars, researchers, and students.¹ It intends to provide “effective preservation” of the archived content and provides end user access to content in a limited set of situations (see Section 3.3.2.4, *Usability of Information*)

¹ <http://www.portico.org/index.html>

Publishers

The benefits to publishers include reducing (or eliminating) the publisher's internal archiving costs, meeting library demand for a trusted, third-party archive, and meeting library demand for perpetual access without negative impact on publisher's operations.

Libraries

The benefits to libraries include securing protection against eventual loss of access to important scholarly source materials and providing a practical way to maintain continuity of library collections.

The Portico service to libraries does not alleviate libraries' need for access to new or current scholarly electronic journals. Instead, Portico enables libraries to take advantage of a system-wide approach to preservation of electronic journals, and avoid the costs of locally archiving and maintaining electronic journal literature. Portico may also enable libraries to reduce or eliminate their need for print subscriptions while ensuring continuity of access to scholarly literature.

3.1.3 Organizational Structure (Brief Overview)

Portico is a service that operates under the auspices of Ithaka Harbors, Inc. Ithaka Harbors (Ithaka) is a non-profit organization with a mission to "accelerate the productive uses of information technologies for the benefit of higher education around the world" through shared services. As an "incubated entity" of Ithaka, Portico still operates in the shared services environment under which it was developed and launched. It has five different functional units including Finance, Publisher Relations, Library Relations, Operations, and Technology, as well as the Executive Director's Office. Of these units, several are managed and operated by Ithaka, including Finance, some of the Technology Unit, as well as the Human Resources and Legal Services units which fall within the Executive Director's Office. A detailed view of Portico, as well as the functions provide by Ithaka can be seen in Appendix 4.4, *Portico Functional Overview*.

3.1.4 Technical Architecture (Brief Overview)

The technical architecture of the Portico Archiving service was specifically designed for e-journal archiving, with a logical and physical separation of service & delivery from the archived content. The archiving portion of the system utilizes a content preparation system (ConPrep) as a key workflow and system. It accepts proprietary publisher submissions, determines content, normalizes formats, gathers appropriate metadata and packages it in a standardized way for long-term archiving. The delivery system is completely separate and is managed by JSTOR under a service agreement with Portico. The delivery system will use a modified version of the well-established JSTOR delivery system, including a new interface designed for Portico. More information about this can be found in Section 3.3.3, *Technical Analysis*.

3.2 Objectives, Scope, & Methodology

3.2.1 Scope

This audit evaluated and will provide information on the following topics:

- Organizational Infrastructure
- Technical Analysis (Digital Object Management, Technologies and Technical Infrastructure)
- Content
- Vulnerabilities

In all areas, the focus was on identifying and describing issues that could affect the viability and stability of the repository and the digital objects stored within the system.

At the time of the audit, Portico was just moving from beta to production versions of the archiving system and content had yet to be ingested therefore viewing and testing content was out of scope for this audit. The access system – also still under development – was viewed in its beta version during the onsite visit and was later viewed through the delivery system provided to libraries and publishers for verification purposes (temporary username/password logons provided by Eileen Fenton). Auditors were provided complete access to developers and system personnel where needed in order to discuss and facilitate analysis of the Portico system, its development, issues discovered to date, as well as information about loading expectations and other imminent work.

3.2.2 Method of Work

The work performed in this audit consisted of a review of documentation (both publicly available information and sets of questions provided in advance); interviews conducted with key staff; an onsite visit; completion of the RLG-NARA Checklist for the Certification of Digital Repositories; and observations. The onsite visit included inspection of server rooms, network utilities, HVAC provisions, and servers associated with providing information to users. Security arrangements for the server facility on the Princeton campus were also observed and questioned. No detailed technical testing was conducted although technical auditors created several scenarios to which technical staff had to verbally discuss and respond. These scenarios were designed to detect potential vulnerabilities in policies, functionality (ingest, processing, archival package creation, data loss detection & resolution, access, etc), staff knowledge, and facilities. Such scenario testing cannot test the reliability of the digital records within the digital archive, but can provide insight into threat detection and risk management capabilities. Finally, an analysis of content was made to investigate discovery and delivery options to be made available should the Portico archive be made accessible due to a trigger event.

3.2.3 Standards Against Which Audit Was Completed

The RLG-NARA Checklist for the Certification of Digital Repositories (August 2005) provided the metrics for this audit. The checklist was developed by an international task force of experts in digital preservation, digital repositories, and data archives. While the Checklist is not yet an international standard itself, it is based upon and references a number of international standards and best practices such as the Reference Model for an Open Archival Information System (ISO 14721:2004), Control Objectives for Information and related Technologies (COBIT) 4.0, Information Technology—Security techniques—Code Of Practice For Information Security Management (BS ISO/IEC 17799:2005), PREMIS Preservation Metadata (2005), and Trusted Digital Repositories: Attributes and Responsibilities (2002).

3.3 Findings

3.3.1 Organizational Analysis

3.3.1.1 *Governance*

Originally called the Electronic-Archiving Initiative, the e-journal archiving concept was begun under the auspices of JSTOR and the Andrew W. Mellon Foundation. By 2005, this activity had migrated to become a part of Ithaka, a new non-profit organization with a mission to “accelerate the productive uses of information technologies for the benefit of higher education around the world.” In 2005, a new electronic archiving service resulting from the Initiative's extensive collaborations with the community was launched under the name Portico and the direction of Eileen Fenton, Executive Director. The new service is establishing itself as an independent organization, but in many ways, retains many of the vestiges of its Ithaka incubation. Ithaka provides many of Portico's operations such as finance, human resources, legal, and IT support through shared services.

Initial support for Portico was provided by JSTOR, The Andrew W. Mellon Foundation, and the Library of Congress. Ithaka continues to incubate and support Portico's development, and the Ithaka Board of Trustees provides formal governance and oversight (See Appendix 4.6). Additional guidance is received from the Portico Advisory Committee (see Appendix 4.5), which is comprised of representatives from the scholarly publishing and academic library communities.

3.3.1.2 *Staff*

The core Portico staff is quite small. At the time of the audit, the staff FTE was approximately 15. This small number is possible because of the use of Shared Services with Ithaka. For example, financial services, human resources, and IT services are shared staff with Ithaka. Many of those staff members are physically located in Ithaka's New York office while some shared IT staff remain on Ithaka's payroll, but physically reside in Portico's Princeton, NJ office. Finally, the entire delivery operations are contracted out to JSTOR and therefore all staff related to delivery systems are on JSTOR's payroll and reside in JSTOR's Ann Arbor, MI office. There are benefits and concerns related to this shared services and “contracted out” services model and are expressed elsewhere in this audit report.

Current staffing levels were deemed adequate at the time although it was anticipated that with new agreements from large publishers, there may need to be some staffing increases within the Portico Technology Unit in order to manage the dramatically increased workload. Fenton and Ithaka have expressed the opinion that “explosive growth is bad, controlled growth is desirable” so the plans are to ramp up slowly and manage backlogs as necessary. They do have a Content Capacity Planning Group and their focus has been on the first three years (2006/07/08) with “reasonable” targets. Targets were set in mid-05 and need to be reassessed. These reassessments are planned to take place on a six-month cycle.

Job descriptions were asserted to be accurately reflecting expectations and actual responsibilities -- at least as much as they could in a young organization that is evolving daily. In all interactions, Portico staff and affiliated Ithaka and JSTOR staff appeared to the auditors very knowledgeable and capable.

3.3.1.3 *Policies and Procedures*

Policy and procedure documentation is not robust. On the one hand, the evolution of functional requirements, specifications, and agreements can be traced through a regimented series of drafts and versions. On the other hand, other kinds of documentation are lacking at this stage. The latter is likely to stem from the relative youth of the organization, as well as the “just do it” spirit of the staff. Portico is

developing and building a new kind of system from the ground up and it appears that at this moment development and production schedules take precedent over the documentation of policies and procedures.

Portico strives to provide transparency in its operations. Through existing documentation, they can provide information about changes to its operations, procedures, software, and hardware as all of these are tracked through an evolution of technical specifications. Service model changes, a shift in Portico's pricing structures for libraries and publishers for instance, are documented though these. In reality, the technical infrastructure was still in true development at the time of the audit and many of the decisions are documented in internal email and design discussions. Once they achieve full production capacity, Portico plans to begin more formal documentation of this though in the interim, information about these potentially important change management decisions is not easily accessible.

Portico has also committed to defining, collecting, tracking, and providing on demand its information integrity measurements. They have begun to define a set of metrics that could be shared (transparent to "outsiders"). Portico will produce reports on request for publishers and eventually to user/subscribers. Contracts with publishers are changing so that publishers have a mechanism to ask for statistics (problem reports, quantity of material/titles/articles processed in a month, etc.).

This is clearly an area where more work is needed, but as mentioned above, the current deficiencies appear to be less purposeful than a function of rapid development of the system so that a production environment could be deployed and archiving enabled.

3.3.1.4 *Financial Analysis*

Because Portico is not a separate legal entity, but rather is a unit within the Ithaka Harbors, Inc., organization, audited financial statements for its operations were not available. Our financial analysis draws from three sources: audited 2004 financial statements of the parent organization, Ithaka Harbors, Inc.; Portico statement of activities for 2004 (audited) and 2005 (unaudited); and Portico's response to Advance Technical Questions (Appendix 3.3). Initial funding for Portico in the amount of \$6,363,708, was provided by JSTOR, Ithaka, and The Andrew W. Mellon Foundation. In addition, Portico was recently awarded a grant for \$3 million by the Library of Congress as part of the National Digital Information Infrastructure and Preservation Program (NDIIPP) to support further development of the Portico organizational and technological infrastructure, and an economic model capable of sustaining a long-term preservation archiving service.

Portico's revenues for 2006 are projected to be as follows:

Support from Libraries	\$1,089,767
Support from Publishers	223,250
Foundation and Public Support	3,100,000
Interest & Dividends	25,000
Ithaka Support (Internal)	975,000
<u>Total</u>	<u>\$5,413,017</u>

Expenses for the same year are expected to total \$5,290,188, producing a net surplus for the year of \$122,829. At this stage Portico relies most heavily on grant funding to meet its expenses, although it project a gradual tapering off of this reliance over the next several years.

As an archiving service with a long-term perspective, Portico believes it is important to adopt an economic model which permits the archive to build diversified sources of revenues which can be used to cover ongoing operating costs. Support for the archive should come from the primary beneficiaries of the

service - libraries and publishers - but support from other sources, such as charitable foundations and government agencies, is also expected. Portico's community-based approach rests upon the assumption that libraries and publishers will share the costs of the electronic archiving service and will secure the benefits and opportunities for savings that the archive enables.

Business Model

Portico's business plan seems to be well-considered and sound, with some reservations. Portico's aim is to provide services to publishers and libraries by serving as a centralized repository. Publishers who participate are able to submit their entire title list for archiving in Portico along with a relatively small amount of money and in return, Portico provides publishers with the ability to concentrate on the core business of publishing rather than long-term digital preservation. The incentive for libraries to invest in Portico is the ability to exploit its shared preservation and storage capabilities, alleviating the need to preserve the publications on a local level. According to Portico, the service provides libraries "a practical means to act upon their traditional preservation mandate" by investing in the centralized service. It also provides libraries a mechanism for promised "perpetual access" to e-journals, should the publisher designate Portico as the provider to do so.

Revenues from publishers are expected to be modest. The true goal is to encourage publishers to deposit their content rather than have Portico recoup all costs. Participating publishers must pay a "Supporting Publisher Contribution" in addition to depositing their content. It is not clear why this is termed a "contribution" rather than a fee. This tiered fee is based on self-reported annual journal revenues and range from \$250 to \$75,000. This revenue is designed to fund initial conversion tool development and to cover the cost of new content as it is published.

Revenues from libraries are far more significant and represent 75-85% of the planned revenue. Since libraries are seen as the primary beneficiaries of Portico's service, they must provide the majority of the revenue. To participate, libraries must pay the Annual Archive Support payment, a fee based on self-reported library materials expenditures or LME (which is essentially the content Portico is created to insure). Library support contributions are roughly equivalent to 1% of an institution's LME, making fees range from \$750 up to \$24,000.

Additionally, all libraries that support (subscribe to) Portico in 2006 or 2007 will be designated "Portico Archive Founders." The 2006 signers of this five-year commitment will receive a 25% savings each year of the five years. Libraries signing on in 2007 will receive a 10% discount for each of the five years of their commitment.

One apparent drawback of this fee structure is the seeming lack of a correlation between the revenue sources and three important cost drivers for Portico: the amount of journal content managed; the complexity of that content; and the eventuality of having to make that content available to a large audience. The publisher's annual contribution to Portico is based on the publisher's (self-reported) revenues, rather than the number of titles or type of literature archived. The revenue from libraries is based upon each library's annual materials expenditures, and is thus only indirectly related to the amount of content Portico holds on the library's behalf. Hence it is not clear how future migrations will be financed, although Portico has made provisions to minimize those costs by normalizing the content it maintains and by automating many of its processes.

Revenue & Expenses

As a relatively new service, Portico's 2004-2005 budget years – funded entirely by grants and support from other organizations – cannot be considered to reflect future budgets and financial activity, unless Portico is able to maintain its track record of sizable grant monies received. At the same time, the budgets can be used as indicators of fiscal stability and solid financial management.

Audited statement of Portico activities from 2004 show a strong net surplus on the year as a result of over \$3 million in support from the Mellon Foundation and Ithaka. Unaudited numbers for 2005 (provided at the time of audit) reveal a much smaller surplus – evidence that Portico made sizable investments in technology, public service, and various outreach efforts in order to establish the archiving service. The year ended with a net surplus of approximately \$482,000 despite increased revenue over the previous fiscal year.

At the audit, Portico’s executive director provided a copy of the 2006 budget for Portico. This budget is the first one to include anticipated revenue from publishers and libraries, as well as \$3 million award from the National Digital Information Infrastructure and Preservation Program (NDIIPP) and continued strong funding from Ithaka. Still, this first year involving a move from development to production is anticipated to have the greatest amount of expenses, with sizable expenses in capital technology acquisitions as well as operations support.

As Portico moves forward, ongoing costs of the archive including repository costs, capital improvements, enhancements, and providing access in the event of publisher failure are expected to follow the overall cost distribution pattern of the Portico economic model: approximately 5-10% from publishers, 10-15% from government agencies and charitable foundations, and 75-85% from libraries.

Budget Direction & Cost Controls

Budget and cost controls are managed at a variety of levels. Portico is still operating under the auspices of Ithaka and many of the services, including financial management, are provided through the shared services contract with Ithaka. Portico’s executive director is also responsible for the budgetary sustainability of the organization and actively works with the Ithaka Board of Trustees on issues of strategic directions and business planning. Input and assistance has proven to be valuable because the business model for Portico has had several different iterations before reaching its current one. The latest model seems to more adequately distribute costs against benefits gained from the service and while libraries certainly want to see publishers increase their monetary contributions, Portico seems to have chosen correctly in emphasizing content acquisition over full financial parity in the early years.

There is some concern about the interrelated organizations (Portico and Ithaka) partnering on services, as well as funding. It’s often difficult to determine with precision Portico’s actual costs because it appears that funding is a bit circular (revenue –grants—from Ithaka to Portico but then Ithaka services as expenses). This should be better clarified so funding sustainability can be better understood.

3.3.1.5 *Contracts (Submission Agreements) & Licenses*

Critical to Portico’s services are their contracts and license agreements. These fall into three main categories: publishers, libraries, and shared Ithaka services. The first two are most important when considering long-term access to the intellectual content of the e-journals in Portico’s care. The latter is important for the day-to-day operations and management of Portico.

Publisher Agreements

Publishers who participate in Portico sign a Publication License Agreement, granting perpetual non-exclusive rights to Ithaka Harbors, Portico’s parent organization, according to the business model discussed above. The term of each license is three years. At the publisher’s option, Portico may provide post-cancellation “perpetual access” to Portico supporting libraries on behalf of the publisher in conformance with their customer agreements. Publishers deposit content in a timely way following publication and make an annual financial contribution to support the archive.

In the event that a publisher decides to terminate the agreement, the publisher no longer has to pay the annual publisher contribution or deposit new content. In return, Portico maintains the right to continue to “retain, migrate, and deliver copies” of the archival versions of publications already in the system at the time of termination. This means that Portico can truly offer “perpetual and irrevocable access” to e-journal titles, even if a publisher terminates its work with Portico. This is a strong benefit for libraries and is likely to serve as a powerful incentive.

Library Agreements

Libraries who participate in Portico sign a Journal Archive License Agreement and make a one-time contribution followed by annual financial payments. In return, Portico provides the following to libraries:

- bibliographic and preservation-related information about titles and issues within the system
- limited access (up to 4 staff) to Portico content for purposes of verification and testing;
- perpetual access to e-journals to which an institution may have previously subscribed and entered into a perpetual access agreement with the publisher (possible only if publisher designates these rights to Portico); and
- access to publisher content (all titles belonging to publisher) in Portico if a trigger event occurs and when the titles are no longer available from the publisher or other sources.

Library agreements are five years in duration.

Ithaka Shared Services

This set of contractual obligations and services is by far more critical to Portico in many ways. Because Portico was “incubated” by Ithaka, its services were developed and remain embedded in the Ithaka Shared Services model. In many ways, these shared services allow the core Portico staff to concentrate on the core Portico services – archiving e-journals. At the same time, there seems to be some very interesting ties between Ithaka and Portico which are often difficult to unravel and fully understand. See comment in Section 3.3.1.4, *Financial Analysis, Budget and Cost Controls* for more information.

3.3.1.6 Succession planning

Should Portico fail, they anticipate passing on their archival obligations over to another not-for-profit organization. In their publisher agreement, they have secured the rights necessary to do so. Commitments to do so are also made in the library agreement.

They are in very preliminary discussions with possible succession candidates.

3.3.2 Content Analysis

3.3.2.1 *Logical and Physical Content*

The Portico approach to archiving electronic journals is based on archiving “source files” rather than presentation or delivery files. Source files are electronic files containing graphics, text, or other material that comprise an electronic journal article, issue, or volume. Source files may differ from files presented by the publishers online most typically by including more information or higher quality graphics. Publishers deliver to Portico the source files of their electronic journal editions (e.g., SGML or XML text; PDF or PostScript page images; graphics, media, and supplemental files); Portico normalizes the text files from their original proprietary formats to an XML DTD based on the National Library of Medicine Archival DTD and deposits the content in the Portico repository. Portico retains the deposited files for the long term through bit preservation within its Archival Information package and in doing so supports critical authenticity verification for the normalized content. The normalized files are both a preservation strategy as well as the target of preservation and will be migrated as needed to new formats and technologies. Portico preserves the intellectual content of the journal, including the text, images, and functionality such as internal and external linking, when included in the provided data. The exact “look and feel” of any HTML rendition is not targeted for preservation and the value-added features embedded in publisher delivery systems such as e-commerce functions and personalization are clearly not part of the preservable content provided by publishers and therefore are not replicated within the Portico system..

3.3.2.2 *Extent (titles, date ranges, etc)*

Committed Content

As of 22 June 2006, Portico had 13 publishers “committed” to contracts with them, representing 3,557 electronic journal titles.² Portico agreements stipulate that the publisher intends to deliver to Portico electronic files of back issues of the electronic journals “to the extent [they are] available” to the publisher. Thus one anticipates that all issues of these titles will eventually be available in Portico. Delays in availability would be related to maximum loading capabilities and the extent to which complete title runs are currently available. See Appendix 4.10, *Portico Participants Meeting at ALA Handout* for more information about publisher and title commitments.

Loaded Content

As a young archive that was still in the development process and moving toward a production environment at the time of audit, Portico’s loaded content does not yet match the “committed content” from publishers. In all likelihood, it will take up several years to load the complete content from the committed title list (contemporary, “born digital” content). Additional content in the form of digitized backruns of journal titles will be made available through publisher digitization and will be added to Portico as available. This places additional materials into the loading queue and will demand Portico prioritize their load list. The backlog isn’t concerning since the Portico business model is based on the concept of a dim preservation archive (like “insurance”) rather than a day-to-day database content provider. Still, the load/ingest expectations of Portico and timing of content availability should be made available to supporting libraries if not via the website.

As of 22 June 2006 however, over 20,000 pages had been ingested, from committed publications from the American Mathematical Society, Berkeley Electronic Press, John Wiley & Sons, and the Oxford

² Current list of committed titles available on Portico website, http://www.portico.org/about/committed_titles_alpha.html

University Press.³ While certainly an accomplishment, realizing that the content of the committed journals actually represents millions of pages to be ingested, it is clear that Portico will be ingesting content for many years before it will be able to fully provide all committed titles in the event of a trigger event.

3.3.2.3 Usability of the Information

Portico's mission, to ensure that electronic journals remain available to future generations of scholars, researchers, and students, clearly identifies Portico's ultimate user community as that set of individuals who at some point in the future must turn to the archive for access to importantly scholarly resources. To meet that mission, Portico has chosen to base future usability on an upfront preservation strategy: format normalization.

By normalizing the textual data of the e-journal articles into a standardized, commonly used XML DTD, Portico is able to uniformly manage the data, control presentation, enable searching, and maintain other functionality which would otherwise be complicated, time-consuming, and probably prohibitive, because of the varied and proprietary formats deposited by publishers. As "flat" encoded text, migrations to new formats and technologies can be accomplished as needed.

Although the preservation strategy utilizes the NLM Archival DTD, Portico also maintains the deposited files for the long-term. In doing so, they can not only trace authenticity, but could conceivably go back to the original file should that be necessary. It is conceivable that these bit-preserved original files would need some kind of preservation action to render them if necessary at some point in the future, but since many of the publishers are using fairly well-documented file formats as a part of their production, it is likely that solutions for these kind of common files will be available through file format registries and supporting services. It is important to note that Portico's philosophy is to preserve the intellectual content of the journal, namely the text, images, and such functionality as internal and external linking (when included in the provided data). *The exact "look and feel" of the HTML rendition on the publisher website and the publishers' value-add features available through their respective delivery systems such as e-commerce functions and personalization are not part of the actual journal content and therefore are not delivered to or preserved within Portico.*

These differences were apparent when viewing Portico content through their delivery system versus viewing the same content as offered directly by the publisher. Auditors were provided access to the delivery system that is provided to staff of libraries and publishers for the purposes of verification and testing and compared that content to that offered up through two campuses subscription packages. While there are obvious differences in content appearance and delivery system functionality, these were anticipated. Portico has made great efforts to articulate exactly what content is being preserved and made available through their system. Auditing the content delivery on those principles, the content viewed through the Portico delivery system is very usable and met auditor expectations for usability.

3.3.2.4 Content Accessibility

The Portico Publication License Agreement reserves for Portico a liberal set of rights to use publisher content. Under this agreement accessibility is related to the three "classes" of users served by Portico: publishers, libraries, and the library researchers. Because Portico is not an active (or "light") archive, day-to-day use of the content is prohibited. Rather, Portico follows a "dim" approach to preservation: wide access to archived content is provided only when access to that content is no longer "actively supported" by the publisher. Therefore, access to Portico content is only permitted under controlled

³ Conversation with Evan Owens, 24 June 2006.

conditions and in a limited number of circumstances: for verification, to fulfill a perpetual access agreement, or when a specified “trigger event” occurs.

Publishers and librarians at subscribing institutions are provided limited, secure access to the Portico service, but only as a verification mechanism. Portico license agreements specify that “Portico will provide access to the Archival Versions for the purposes of verification and testing...Licensee may designate up to four staff members per campus covered in [the] agreement with authority to verify or test the integrity of Portico’s archive via password-protected or otherwise secure access to Archival Versions.”⁴ This limited access verification feature is applicable for contributing publishers and subscribing libraries, and provides a useful measure of auditability for the archive.

Scholarly use of the resources archived in Portico is only possible in two situations: when campus-wide access is opened because of a trigger event or when a perpetual access agreement is activated.

Trigger Events

Campus-wide access for participating institutions will be triggered when any of the following occur:

- (A) **The Publisher is No Longer in Business.** Licensor is no longer in business or is no longer in the business of publishing or providing access to previously published issues of scholarly journals.
- (B) **Title No Longer Offered.** Licensor has stopped publishing and is no longer providing access to the Publication and its back issues.
- (C) **Back Issues No Longer Available.** Licensor has stopped offering or providing access to some or all of the back issues of the Publication for a period longer than ninety (90) days.
- (D) **Catastrophic Failure.** Licensor has stopped publishing or providing access to the Publication for a period longer than ninety (90) days due to technical difficulties or any business interruption, bankruptcy, insolvency, receivership or business failure.⁵

In some instances a publisher may also choose to grant Portico the right to provide access to its archived journal titles to meet the publisher’s perpetual access obligations.

According to the license, if any of the trigger events occur and Portico is in operation, Portico will provide access to the applicable archived content for its authorized users until the publication is once again offered by the publisher or a successor. A publisher may also request that Portico provide access to subscribers on its behalf in the event of an interruption in service.

Triggered Content

An interesting choice is Portico’s option to open all of a publisher’s title list to a subscribing institution if a trigger event occurs involving that particular publisher’s content. For example, Institution X subscribes to Portico and also has a yearly subscription to access 30 of Publisher Y’s titles via the publisher website. Publisher Y’s system experiences a catastrophic failure. Should one of the specified trigger events occur, once Portico follows the proper procedure, Institution X would gain access to all of Publisher Y’s titles deposited in Portico, not just the 30 to which they originally subscribed.

Availability (Timing) of Triggered Content

Trigger events do not trigger immediate access via Portico. While Portico has the functional capabilities to merely “flip the switch” to provide access, licenses specify that Portico must give written notice to the

⁴ *Portico Journal Archive License Agreement*. Ithaka Harbors, Inc., 2005-2006

⁵ *Ibid.*

publisher about its intent to provide wide access to any applicable archival versions “after an agreed upon period not to exceed sixty days.” This means that subscribing institutions could conceivably be without subscribed resources for up to sixty days before Portico can legally light its archive.

Perpetual Access

Some publisher licenses guarantee former purchasers and subscribers continued access to back issues of titles even after their subscriptions to those titles cease. This access is generally referred to as “perpetual access.” Although some publishers manage perpetual access through their own local mechanisms, some publishers have designated Portico as the provider of perpetual access services for their titles. (See Appendix 4.8, *Portico Participants Meeting at ALA (June 2006) Handout*, for the list of publishers who have designated Portico to provide perpetual access.)

3.3.2.5 *Link management solutions*

Every “content unit” within Portico (typically, a content unit refers to an e-journal article) has a unique Portico archival identifier. It is searchable, but not otherwise actionable.⁶

Some publications may also have a Digital Object Identifier (DOI). The DOI is increasingly supported within the publishing industry and while not required for Portico deposit, if a publication arrives with a DOI, the DOI will be maintained within the curated metadata section of the Portico METS package representing the archival version. These DOIs could potentially be used for external or internal linking to other articles within references, but this feature is not currently implemented in Portico.

3.3.2.6 *Termination of Service policy*

Portico has two license agreement issues related to termination of service: an institution’s termination of their agreement with Portico and the potential cessation of Portico itself.

In the event that an institution terminates its agreement with Portico, all access of staff and authorized users would be terminated. The only exception for continued use would be if an institution activated a perpetual access option and Portico was designated as the publisher’s provider of that service. In the event of that situation, the institution would gain access to publications only covered by the perpetual access agreement.

In the unlikely event that Portico ceases to operate, the license agreements with libraries and publishers stipulate that a successor non-profit organization would be identified to maintain the archival versions. In the event that a successor cannot be found, Portico would provide “a copy of the Archival Versions available to Portico to an appropriate not-for-profit organization and assign its rights to that institution.”⁷ Portico has thus both committed itself, and contractually secured the rights, to perpetuate its preservation of its archived journal content.

3.3.2.7 *Threats to Content*

Threats to networked digital archives are potentially similar to those faced by other content in electronic databases that are connected to networks. The system may be vulnerable to loss of content through any of the following threats:

- hackers/malicious data change
- technical hardware failure
- software error introduction
- bit loss

⁶ *Portico Delivery Functional Requirements*, version 1.3, November 11, 2005.

⁷ *Portico Journal License Agreement*, clause 6.4.

- environmental issues (loss of electricity, failure of cooling systems, small-scale flooding, etc.)
- large disasters (either localized or regional).

To mitigate these risks, it is important that they are assessed and that the policies and procedures, as well as the technical systems employed by Portico, address these threats or vulnerabilities.

3.3.3 Technical Analysis

3.3.3.1 *Architecture*

The Portico hardware and software systems were designed by the Portico Technology Unit, currently a staff of eleven headed by Evan Owens, Chief Technology Officer. Infrastructure services (networking, hardware installation, configuration, and administration, and database administration) are provided by the Shared IT unit of Ithaka. (A component of Ithaka's mission is "providing administrative and technological services that can be shared by affiliated entities to increase effectiveness, lower costs and allow affiliates' to focus efforts on mission-related activities."⁸)

The resulting technical infrastructure is developed and managed by a combination of Portico and Ithaka staff and services supplied by Princeton University. The systems hardware and masters or primary copies of all electronic journals are stored in the server room of the Princeton University Computing Center. Network connectivity for the system is also contracted through Princeton University. Local server management (hardware, software, and backups) is performed by the Ithaka Shared IT staff on behalf of Portico. Finally, access management is provided through JSTOR, leveraging its existing extensive infrastructure for delivery. Portico will export data from the archive system on a regular basis to the delivery platform and system managed by JSTOR. All components of the archive system are operational and the delivery system was being implemented at the time of the audit.

The solution architecture is divided into four sub-processes:

- Provider Setup – R&D of provider content, which includes tool development and testing, both before and after agreement (which is then tested in a QA and Staging environment pre-production)
- Content Preparation – Automated processing of content in preparation for archiving
- Archive – Replicated storage of provider content and metadata
- Delivery – Distribution to content consumers via JSTOR

ConPrep is designed such that it consists of several components, or runtime processes. A key architectural property of these components is that they can be distributed across physical machines in any manner.

⁸ Ithaka, Mission and background. Retrieved Aug. 02, 2005, from <http://www.ithaka.org/about/mission.htm>.

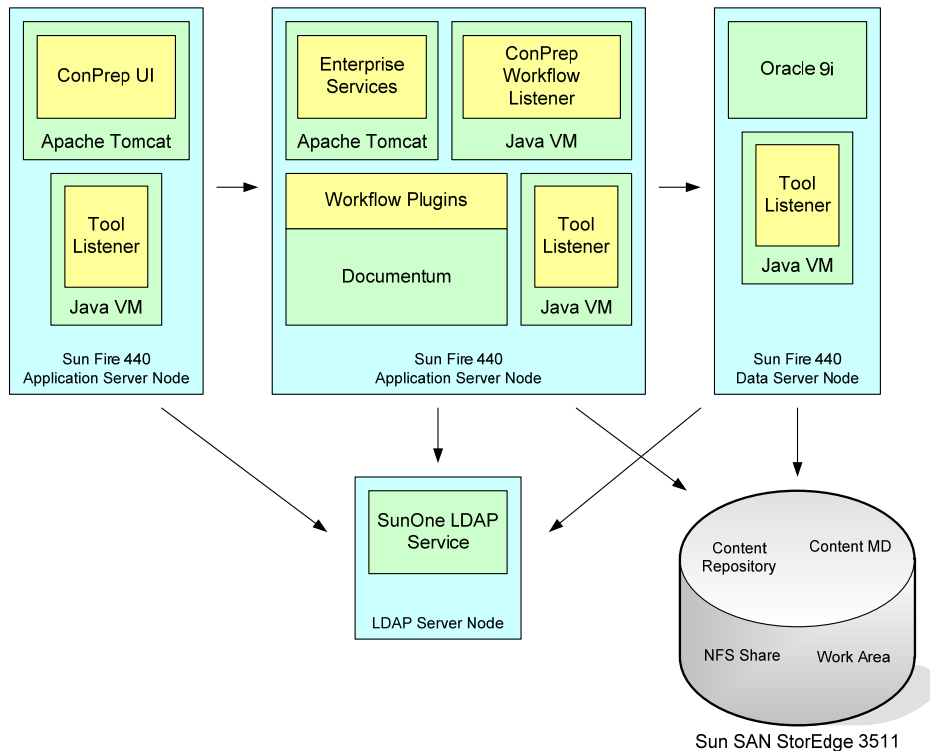


Figure 1 – ConPrep System Architecture, Deployment View⁹

Documentum is an Enterprise Content Management (ECM) system originally developed by Xerox and now owned by EMC. It is a proprietary system, but because of its mass deployment (in use in government agencies, most major financial institutions and insurance companies, as well as corporations with mandatory content retention compliance issues such as Boeing and McDonald Douglas), there is a very low probability of system abandonment or loss of support.¹⁰ Content Preparation (ConPrep) employs Documentum as a platform for both management of provider content and workflow processing. Documentum stores all assets and metadata as objects, accessible via a programmatic interface and managed via the WebTop User Interface. Documentum objects, depending on type, can contain content (as in the case of a file), contain other objects (as in the case of a folder), hold attributes, be archived, be versioned, or any combination of the above. The structure map for Portico e-journal issues, for example, is constructed with Documentum objects.

Architecture Options Considered

Portico evaluated other systems, including DSpace and Fedora (and does not rule out the possibility of using Fedora at a later date) but feels Documentum provides appropriate functionality. Portico has focused on ensuring that there are not too many dependencies on any particular technology. They rely on Documentum for only about a dozen Java classes out of several hundred; except for the user interface, there is very little dependency on Documentum.

3.3.3.2 Data Security (Access controls)

In the context of Portico architecture, authorization is distinct from authentication. In addition, authorization is extended via action control lists, per user. Another important distinction is that the

⁹ *Portico Content Preparation: System Architecture*, version 1.0, 17 February 2006.

¹⁰ *What is Documentum?* WikiD4D, http://www.wikid4d.com/wiki/index.php/What_is_Documentum%3F%3F%3F

archive management application is separate and distinct from the delivery application (i.e., JSTOR); there is no public (non-staff) access to systems.

Authentication

Authentication is based on Documentum's authentication subsystem which defines users either through LDAP or UNIX accounts. Current users in the system are defined using LDAP (at the time of the visit, Portico was not sure of which version was used, noting that the service is offered through shared IT infrastructure). The use of SSL or TLS has been deferred due to the focus of priorities on simple functionality for the initial release, not implementing the encrypted security component in the initial design. All communications with the Archive are behind the corporate firewall. Communication to Documentum is handled through its proprietary protocol called DMCL. Communications between tools are implemented using JMS. Configuration information including accounts and passwords are stored on an LDAP server. The Documentum Content Distribution Service, which handles distribution of content to Portico's external delivery system, uses the ICE protocol over http. Distribution over ICE is entirely behind the company firewall.

Authorization

Authorization is controlled by native Documentum functionality, based on LDAP. Delivery is a separate system controlled by JSTOR and applies only to the service copy of the Portico archive managed by JSTOR. Within the JSTOR system, an external application server handles all authentication and authorization for the delivery site, which store user accounts, sites, and roles in its own database. Some of the authorization data is created in that system programmatically, as data is processed, and some is created manually. The granularity of users and groups is quite flexible – the business access model allows most authenticated users to be authorized for all data in the system. Some users are only authorized for some data.

Physical Security

Portico's servers reside in the Princeton University data center. These are secure facilities. A limited number of Ithaka Shared IT staff have entry, via finger print recognition. Also, the Ithaka firewall, which Ithaka fully controls, resides in the Princeton secure facility. No access logs are routinely shared with Portico staff so that they could ascertain the safety of their servers. Portico would have to request access logs for the physical facility if necessary. This is not an opportune situation for managing the absolute security of Portico, Ithaka, and JSTOR servers and data. Appropriate and regular logs should be delivered and monitored by Ithaka IT staff to better ensure machine and data security.

3.3.3.3 Data Deposit & Ingest

General

Electronic journal content ingested into Portico is determined by agreements with publishers. Agreements specify titles, extent (default is entire runs unless otherwise specified in the agreement), as well as the kinds of metadata the publisher will need to provide with the content. Further documentation is exchanged with the publisher prior to deposit in order for Portico to understand the kinds of source files and packaging that comprise "publications" coming from publishers.

Because publishers use a variety of content management systems, Portico cannot expect "standardized" packages of content. Instead, Portico must understand the issues related to publisher production – often different for different titles from same publisher – and then accept the material and create the "standard" package that comprises a journal article or issue. To manage this process, Portico has established a deposit and ingest system which is actually devolved into an automated workflow for ingest. The components of the workflow include [publisher] package disassembly, format identification and

verification, structure mapping, automated metadata harvesting, rule-based format normalization, and support for quality control and inspection. All of these deposit and ingest steps are essentially considered a “pre-processor” workflow that leads to archival ingest and are a part of the Content Preparation System.¹¹

The Content Preparation (ConPrep) System is an essential component of the Portico E-Archiving process. It receives input from content providers and sends its output to the archive system. As such, its design must meet the following goals in order for ConPrep to achieve its role in the Portico solution:

- Provider package interpretation
- Metadata generation and extraction
- Validation, characterization, and normalization of files

The following diagram illustrates the process overview of the ConPrep system:

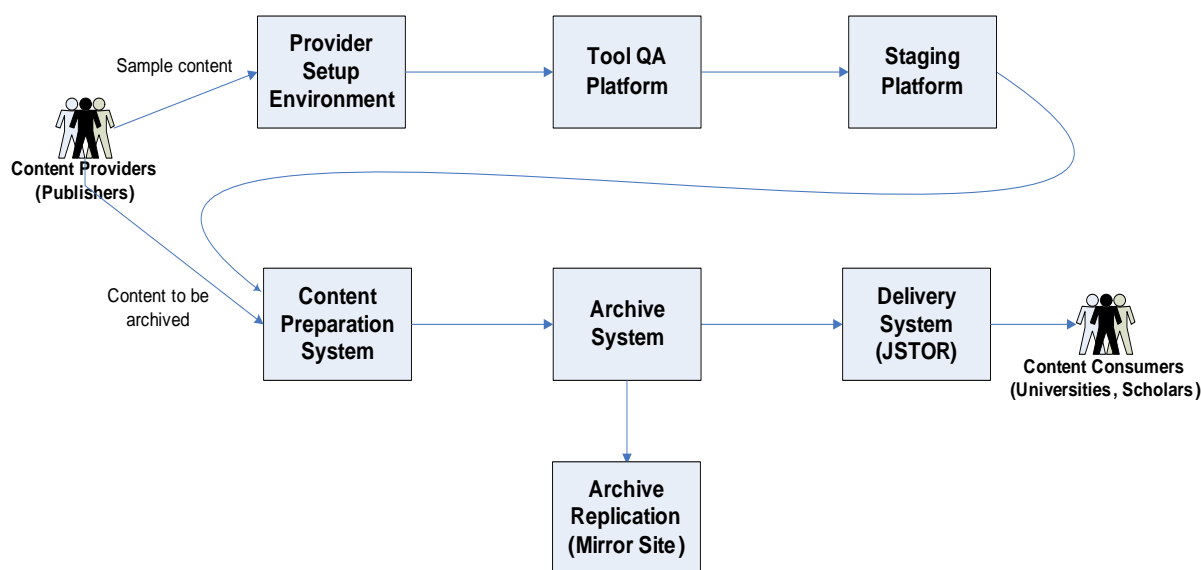


Figure 2 —Portico e-Archiving Process Overview¹²

ConPrep is envisioned as a “pipeline” through which content passes in preparation for archiving. To allow for better understanding of the workflow and its relation to important Portico processes, the following is excerpted from the *Portico Content Preparation System Architecture* document.

The ConPrep “pipeline” has five stages, some of which are automatic and some are manual, and is illustrated in Figure 3:

1. Data is delivered to ConPrep from the content provider then either manually or automatically submitted as a collection of files. This loads the files into Documentum as a batch for workflow processing.
2. Once submitted, the batch must be manually scheduled. This determines whether and when the batch will run.

¹¹ Evan Owens. “Automated Workflow for the Ingest and Preservation of Electronic Journals.” *Proceedings of the IS&T Archiving Conference: Archiving 2006*. Ottawa: August 2006.

¹² *Portico Content Preparation: System Architecture*, version 1.0, 17 February 2006.

3. Once the batch is scheduled, it begins auto-processing in the workflow. The workflow interprets the provider package, extracts metadata, and manages relationships between files.
4. After auto-processing completes, the batch must undergo a Quality Check (QC) to ensure there were no errors and whether it can be released to archive.
5. Finally, a Portico METS file is generated to describe the structure of the content, and the data is released to the archive for long-term storage.

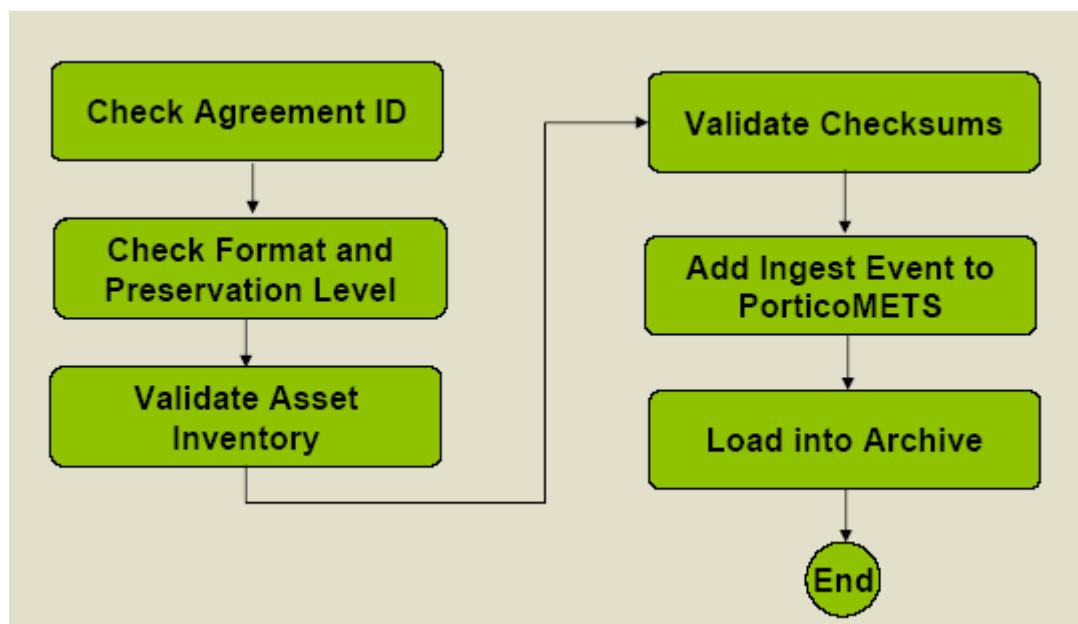


Figure 3 — Archive Ingest Processing¹³

Data Deposit, Source Files & Content Model

Portico acquires its files from publishers through secure electronic deposit. The files comprising publisher content may consist of several source files, each with a different function. They may be SGML, XML, PDF, or one of many types of image formats. Some packages come with an electronic manifest from the publisher while others may not. In order to effectively interpret a provider package, Portico must understand the function of each file and the relationship between them then model them in some meaningful way. Some of the functionality and expectations of what Portico will receive from individual publishers is determined through negotiations and publisher set-up workflows. More general functionality – that is, functionality common to electronic journal content – is addressed and managed through Portico’s content model.

Loosely on MPEG-21 DIDL concepts, the Portico content model consists of the following layers:

- The Content Type
- The Content Set
- The Content Unit
- The Functional Unit
- The Storage Unit

This content model recognizes and functionally replicates the hierarchical model of e-journal content (“e-journal” content type, a particular e-journal (or content set) consists of a series of articles (content units),

¹³ Ibid.

which contains documents, graphics, and media (functional units), which are represented by files (storage units).¹⁴

To maintain these hierarchical relationships in the archive, Portico developed the Portico METS (Metadata Encoding & Transmission Standard) schema to manage the structural relationships and associated metadata with the publisher source content. For Portico, the encapsulation of the structural relationships, metadata and source files yields an “archival record” for the content.¹⁵

File Formats

As noted above, numerous types of file formats are received as a part of publisher submission packages. It is the job of the ConPrep system to derive information about individual file formats for conversion, rendering, and long-term preservation purposes. Using a series of validation tools, the Conprep system determines and records the file types. Information, as needed, is recorded about file types in Portico’s Format Registry. The Format Registry is an XML-based registry consulted to identify, validate, characterize, and render various format instances. The registry is shared by all systems and applications: ConPrep, Archive Management, and Distribution. In the future, the Portico Format Registry could be supplemented by or make references to international format registries under development such as the Global Digital Format Registry (GDFR) or the UK National Archive’s Digital Record Object Identification (DROID) system. Utilizing international, collaboratively built tools such as these will reduce local effort and provide Portico with authoritative format support tools.

Unique Identifiers

Portico must address two kinds of unique identifiers in receiving, ingesting, and managing its data for the long-term. The first is the issue of understanding and managing identifiers issued to source content by the content providers. The second issue is the application of their own unique identifiers so that the content remains unique and identifiable within the Portico archiving system.

According to the Portico ConPrep documentation, a key consideration in the design of the ConPrep system is that each content provider has its own standard for sending data to Portico. Because Portico is a system which aims to process and archive this data in a generalized fashion, a way was devised to capture provider-specific behavior such as file naming conventions, directory hierarchies, and processing characteristics in an externalized fashion. This information must therefore be referenced and/or employed during content preparation. This is the purpose of the Submission Profile, an XML document representing a ConPrep interface agreement with a provider. It contains a name for the profile (which typically has the provider name and version), the type of agreement, a list of file formats expected by this Provider, and a series of pattern matching rules used during processing. These pattern-based rules allow ConPrep to determine how the files are to be organized into units and what kinds of files are to be processed in each stage of the workflow. When a new provider intends to send content to Portico for archiving, the provider must deliver sample data to Portico for testing purposes and a new Submission Profile must be constructed to represent the interface agreement.

For its own unique identifier system, Portico uses the Archival Resource Key (ARK) system developed by John Kunze of the California Digital Library (CDL). The Archival Resource Key (ARK) identifier is a naming scheme for persistent access to digital objects (including images, texts, data sets, and finding aids).¹⁶ [ARK is currently under consideration by the Internet Engineering Task Force (IETF) and

¹⁴ Ibid.

¹⁵ According to Evan Owens, the Portico METS schema is loosely based on version 1.4 of the [official METS standard](#). It is not strictly speaking a “profile” of METS but a combination of METS and MPEG-21 concepts expressed in METS vocabulary.

¹⁶ Archival Resource Key (ARK), <http://www.cdlib.org/inside/diglib/ark/>

already has registration schemes and Name Assigning Authority Numbers (NAANs). Many digital repositories are already using ARK and consider ARK to be the de facto standard for this.] Portico uses an open source software tool “noid” (Nice Opaque Identifier”) to assign ARKs to all files, archival units, and most metadata blocks. The noid minter tool is a lightweight database designed for efficiently generating, tracking, and binding unique identifiers, and allows Portico to effectively bind important, discrete units of information.¹⁷

Validation

Portico does not currently have a validation tool, but they are working on a Schematron validator.

Metadata Capture

Every object in the archive must have a contract format, identity, table of contents, and descriptive metadata. Objects can have other metadata, but these elements comprise the minimum set. There is also events metadata. There are no automated workflows for contracts yet. Consequently, the METS information is created manually and therefore features less events metadata.

3.3.3.4 Archival Storage

Portico’s archival servers are physically deployed in and managed by the Princeton University Office of Information Technology data center. (Delivery systems are completely separate and managed by JSTOR.) Everything in the archive is copied to hard media, for protection against “bit rot” of any online material. Two backup copies of the files, on DVDs, are also stored in the data center. In the future, storage of hard media is expected to be supported by a third party vendors (or vendors).¹⁸

Capacities

The storage capacity is separate from the application design in that Documentum can address a variety of storage devices and systems. Portico’s business plans call for approximately 2.5 million pages ingested in 2006 (about 0.5Tb of content). Because of initially slow commitments by publishers, the number of pages ingested by the end of 2006 may be lower. This reflects agreement delays and not technical ingest problems. They expect an annual growth rate of 0.5 Tb to 1Tb for new content, plus publisher backfiles. Portico expects no more than 10TB of content in the first three years, well within the limits of common application software.

3.3.3.5 Preservation Planning (Strategies)

General

One of the key preservation strategies employed by Portico is format normalization. Pre-emptive migration is taking place as a part of submission for publisher-provided DTDs. Upon receipt of data, Portico keeps a copy of the original submission, but for purposes of archiving, creates a version of the publication using an emerging e-journal mark-up standard designed specifically for e-journals. The National Library of Medicine (NLM) Document Type Definition (DTD) is an eXtensible Markup Language (XML) encoding utilized for standardized markup of electronic journals. Using this XML DTD allows Portico to facilitate further rendition of the files in a uniform manner while reducing the number of formats for preservation and future management.

Specific Strategies

Given the relative youth of the Portico archive, there have not yet been any trigger events for preservation activities (e.g., format migration). Considerations for a preservation strategy are in place theoretically –

¹⁷ *NOID Batch Identifier Infrastructure*, <http://www.cdlib.org/inside/diglib/ark/noid.pdf>

¹⁸ *Portico Functional Requirements: Archive*, version 1.0, 14 November 2005.

migration of image formats, conversion of XML files to new XML DTDs, etc. — and can be enacted as the archive evolves. Portico does utilize a series of tools to obtain and document technical information about the files for long-term management and use. Every file in the archive is assigned a format name that points to an entry in the format registry. Every file is also assigned a preservation level: fully supported with promise of migration, supported with reasonable efforts only, or byte-preserved with no promise of migration. The preservation level is determined first from the format validity: a defective file cannot be fully supported; at best Portico can only can promise only reasonable efforts.¹⁹ Further strategies have yet to be determined.

3.3.3.6 Data Management (metadata, logs, etc.)

Management of Archival and Service Files

Designed specifically as a long-term, “dim” archive, the archival versions of the content and the service versions of the content exist in two physically separate environments. The archival data resides on Portico servers, the service versions – periodically copied and exported from the archival versions reside on JSTOR servers as JSTOR is the contracted provider of content delivery.

The archival data is managed in a highly controlled manner – any “touch” of the system must be for strictly specified purposes and must be logged. Backups of the data are contracted to Princeton University OIT. Full backups (file system and database) take place every weekend and incremental backups take place daily.

Metadata

In addition to taking apart a provider package and organizing files into a structure that fits the Portico content model, ConPrep must prepare a fair amount of metadata from the incoming files. There are four distinct kinds of metadata that can be generated:

- Descriptive Metadata
- Technical Metadata
- Rights Metadata
- Events Metadata

Descriptive metadata contains information that describes the content. Since this type of information will vary by the type of content, each content type must have its own metadata format. Article content is specified by PorticoArticleMetadata_1_0 schema, and contains descriptive information about the journal and the provider. This kind of metadata includes ISSN number, volume and issue number, copyright information, article title and date published, etc. While some descriptive metadata is received directly from the provider, others must be derived or extracted from the content.

Technical metadata specifies information about the actual files which contain the content. It includes information such as file size, mime type, format, version, checksums, creation/modification timestamps, etc. For e-journal content, technical metadata is generated by the JHOVE Characterizer tool and hence follows the jhove_1_0 schema.

Rights metadata contains information on the legal rights that are associated with the content. This may include who owns the content, who has rights to use it, in what fashion can it be used, etc. As of Portico release 1.0, rights metadata is not implemented.

¹⁹ Evan Owens. Automated Workflow for the Ingest and Preservation of Electronic Journals. <http://www.portico.org/about/Archiving2006-Owens.pdf>

Event metadata provides an audit trail of decisions that have been made concerning the content. Those decisions could have taken place in either ConPrep or the Archive, automatically by the system or manually by a user. The format of event metadata is specified by the PorticoEventMetaData_1_0 schema. The specifications for the Portico metadata types are each represented as an XML Schema Definition (XSD). Events within Documentum can be logged using Documentum native capabilities, though Portico is not using this functionality yet.

Fixity

The entire batch is first inspected for viruses and ensured that it arrived safely (via a checksum). If an incoming file has a compression or archival layer (such as zip or tar), the layer is removed and all de-layered files are reprocessed. Once all layers are removed, the format of the lead metadata file is verified and transformed (if appropriate), the structure map constructed, and the files are normalized so that the mime type and format is understood. Next, file references (or links) are extracted and resolved. Lastly, metadata is extracted and curated from the batch. The batch is now ready for quality control. Once the batch has undergone manual quality control, the workflow calculates a new checksum for the content, generates and validates a Portico METS, and releases the content to the archive.

Periodically, a process checks the SHA-512 checksums of the files against the ones stored during initial ingest as metadata then produces reports regarding integrity. The actual process of remediation was still open at the time of the audit.

3.3.3.7 Access Management

Access Policy & Delivery Design

By design and license agreement, Portico is a “dim” archive and anticipates limited access to its holdings. In fact, Portico can only provide access to their archived content if a trigger event occurs *and* the publisher is no longer providing the content.

Another component of Portico’s design – this time the technical design – is that the archive is functionally separate from any service capabilities. Portico’s local architecture manages the archival versions only. Provisions for service delivery have been contracted through JSTOR. JSTOR provides the separate, fully-functional delivery of service copies on an as needed basis (publisher/librarian validation is current main use; user delivery in the event of trigger event or perpetual access invocation.).

The following diagram shows the enterprise view of the complete Portico service:

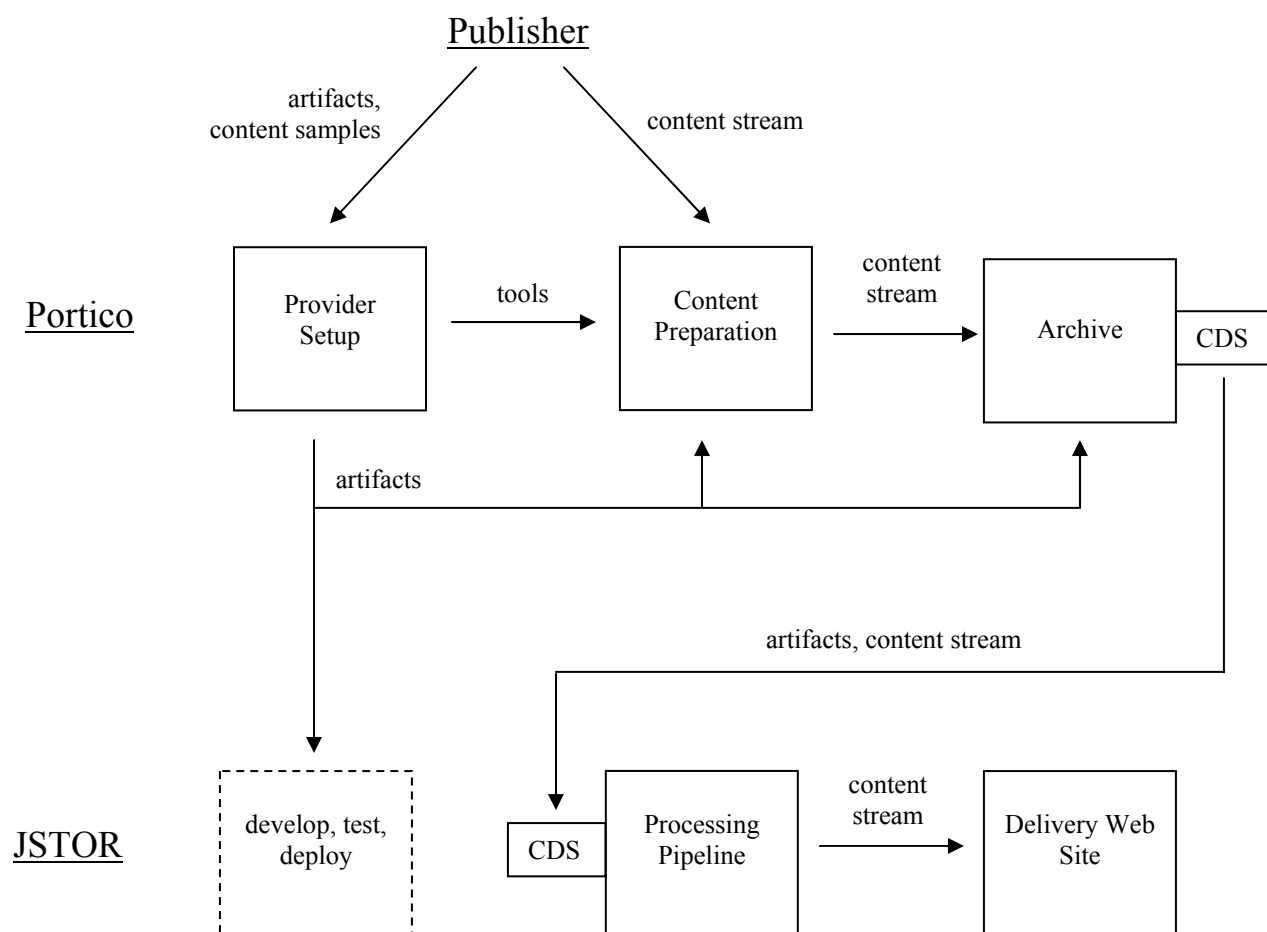


Figure 4 –Enterprise Environment (Publisher, Portico, JSTOR)

Service Levels & Performance Expectations

Delivery: JSTOR

Service level agreements for delivery (i.e., JSTOR) are separate from performance targets for the archive system. The Portico delivery system was designed by JSTOR staff and based upon the long-standing, stable delivery system in place for the JSTOR Journal Storage service. Delivery interface functional requirements, service levels and expectations have been outlined in the *Portico-JSTOR Content Distribution Interface Agreement*.²⁰ A beta version of the delivery system was demonstrated and made available for inspection during the audit and appeared to meet printed specifications.

On the delivery site, browsing and searching that data shall not exceed “normal” browsing speed expectations – sub one-second for most pages on a high speed network connection. Though availability of the delivery site has yet to be a major issue, the intent is to have the site up during normal business hours or better – about 99.9% of the time.

²⁰ *Portico-JSTOR Content Distribution Interface Agreement*, Version 0.2, 18 November 2005.

Archive: Portico

Given that Portico is not available for public access, throughout has been not a primary design concern. Rather, the main priority has been the ability to support regular business operations (e.g., system needs to be up for regular business hours plus any extended hours for maintenance, back-up, etc.).

3.3.3.8 Business Continuity, Environmental Management, and Disaster Planning**Business Continuity**

Portico has not yet developed a formal business continuity plan though components of one exist through a series of service levels agreements (SLAs) with Princeton University Office of Information Technology, Ithaca Shared IT, and JSTOR.

Infrastructure services (networking, hardware installation, configuration, and administration, and database administration) are provided by the Shared IT unit of Ithaca. The systems hardware and masters or primary copies of all electronic journals are stored in the server room of the Princeton University Computing Center. Network connectivity for the system is also contracted through Princeton University. Local server management (hardware, software, and backups) is performed by the Ithaca Shared IT staff on behalf of Portico. Finally, access management is provided through JSTOR, leveraging its existing extensive infrastructure for delivery.

Environmental Management

Portico servers containing the archival content (and current JSTOR delivery system) are housed in the central server room at Princeton University. At the time of the certification and audit visit, it was not clear whether the facility monitored temperature and humidity conditions. (Proof of such capabilities later discerned through inspection of SLA between Princeton and Ithaca Shared IT.) There is no fire suppression at this location which holds many more servers than just the ones belonging to Portico and JSTOR. The facility features a facility-wide UPS with approximately 30 minutes capacity (at least some of the air conditioning is connected to this UPS so it would continue as long as equipment remained running).

All systems are RAID 5 or mirrored, mostly the former. The network patch cables were not labeled at each end, and the racks were not locked. A few (but not many) of the doors on the racks were open at the time of the visit. These are conceivable risks to operations and services should an unauthorized individual gain access to the server room and decide to pull cables. It would be very time consuming to trace network patch cables and restore services – labels on cables would allow recabling much more quickly and reduce potential time off of the network. Media management or migration for these servers has not been an issue to date, given the (relatively new) age of the archive. Policies will need to be developed and processes tested in the near future to ensure that when migration becomes necessary (likely in the next few years), it can occur without problems.

At the Alexander Street computer room, Portico maintains a Windows Active Directory and storage server. These machines manage the in-house desktop network and are not related to Portico archive servers. This facility was below grade, so there is a risk of flooding, and there did not appear to be water monitoring equipment. It lacks a fire suppression system. The door to the facility is locked typically, but it was propped open at the time of the visit to deal with cooling problems (it appeared that the cooling capacity was insufficient for the room). Since Portico offices will be moving from that building soon, only temporary remedies are available or considered for this equipment.

Disaster Planning

As not only a young archiving service, but a young organization, Portico has prioritized system development and functionality over some types of documentation. This includes a disaster plan – a

noticeable absence when addressing risk management. Portico is strongly encouraged to develop a disaster plan which specifically addresses plans for business continuity, as well as environmental disasters of any size.

3.3.4 Vulnerabilities

3.3.4.1 Significant Repository Events

None, thus far. There are plans for moving offices, which might offer an opportunity to design and implement a test restore, media migration or some other preservation activity. It would be important to manage issues such as the security of the backup data during this move.

3.3.4.2 Liabilities

Currently, there is no disaster plan for the repository or recovery, business continuity and organizational survivability in the event of a catastrophic event. Additionally, Portico would benefit from specifically and explicitly articulating the different types of attacks or threat events that might affect the archive.

3.3.5 Observations and Recommendations

Even at this early stage of Portico's development, there are positive signs on many fronts. The service has already shown its nimbleness in adapting its business model to make sure it can achieve its mission in a sustainable way. The precise relationship between Portico, Ithaka, and JSTOR is somewhat ambiguous, and it would serve Portico well to be able to articulate these better to the community they serve. The auditors were given access to variety of information, and Portico has gone to great lengths to publicly disclose information pertinent to its business practices and operations. Yet some confusion and uncertainty remains. This ambiguity is an obstacle to the organizational transparency to which Portico professes to aspire, and which in many other respects it achieves. It is likely to be more of a problem for true outsiders and this can affect the trust of the community in the long-term.

Financial issues may continue to be a concern. Portico seems to be well operated from a budgetary standpoint, but the organization needs to begin to obtain and employ the new revenue sources through publisher and library commitments as the 2006 budget predicts. As mentioned earlier, Portico needs to be able to prove itself to be a sustainable organization to be considered a reliable archive. The pressure will be on in the next few years to get content and subscriber commitments needed to sustain the service. It also remains to be seen how Portico, now a service within the Ithaka organization, will emerge as an independent organization, as planned, and how it will address the new costs, i.e., of systems, legal, accounting, and human resources support that come with that new status.

On the technical front, the auditors saw many positive signs here too. Format choices and technical decisions are solid and well founded, especially for the content types in question. Portico makes excellent use of format registries (and registries in general), reducing redundancies while relying upon authoritative registry information to aid future preservation. There is an independence of the archive from service delivery, which protects the archive from several vulnerabilities. Portico also possesses a good understanding and tracking of events surrounding metadata.

As Portico moves forward, it will be important to document all technical processes, appropriate roles, and decisions. One particularly important aspect of this relates to a disaster plan, which would focus on physical facilities, possible threats or attacks (especially related to possible data loss), and technology infrastructure. The upcoming office move might present an opportunity to design and implement a test restore. Additionally, there is a plan for the archive to be replicated in multiple locations, but this activity has yet started.

Another key area for future consideration is automation. Currently, there are many logs, but few ways to exploit information for consistency checks of the archive. Automation would also reduce the risk of data loss or damage, especially through more routine archive maintenance that does not require human initiation (but would notify individuals in appropriate roles of possible problems).

4 Appendices

- 4.1 Completed Portico Audit Checklist (22 pages)
- 4.2 Portico Responses to Advance Technical Questions
- 4.3 Portico Responses to Advance Financial Questions
- 4.4 Portico Functional Overview
- 4.5 Portico Board of Directors
- 4.6 Ithaka Board of Trustees
- 4.7 Portico Journal Archive License Agreement (Libraries)
- 4.8 Portico Publication License Agreement (Publishers)
- 4.9 Portico Initial Findings Powerpoint Presentation
- 4.10 Portico Participants Meeting at ALA (June 2006) Handout

Appendix 4.1

Completed Portico Audit Checklist

Appendix 4.2

Portico Responses to Advance Technical Questions

Appendix 4.3

Portico Responses to Advance Financial Questions

Appendix 4.4

Portico Functional Overview

Appendix 4.5

Portico Board of Directors

Appendix 4.6

Ithaka Board of Trustees

Appendix 4.7

Portico Journal Archive License Agreement (Libraries)

Appendix 4.8

Portico Publication License Agreement

Appendix 4.9

Portico Initial Findings Powerpoint

Appendix 4.10

Portico Participants Meeting at ALA (June 2006) Handout