White Paper Report

Report ID: 98613

Application Number: PK5007108

Project Director: Eileen Fenton (eileen.fenton@portico.org)

Institution: Ithaka Harbors, Inc.

Reporting Period: 1/1/2009-12/31/2010

Report Due: 3/31/2011

Date Submitted: 4/1/2011

PORTICO

PRESERVATION OF DIGITIZED BOOKS AND OTHER DIGITAL CONTENT HELD BY CULTURAL HERITAGE ORGANIZATIONS

A report for the NEH and IMLS resulting from a grant from the "Advancing Knowledge: The IMLS/NEH Digital Partnership" given to Portico and Cornell University Library

March 2011

A Preservation Model for Cultural Heritage Organizations

TABLE OF CONTENTS

SECTION I. INTRODUCTION AND RESEARCH 4

1. INTRODUCTION 4
2. ARE THE STEPS WE'RE TAKING TO SAFEGUARD OUR CONTENT SUFFICIENT? 6
3. RESEARCH AND ANALYSIS 10
4. OVERVIEW OF CULTURAL HERITAGE ORGANIZATIONS AND THEIR PROJECTS 11
5. THEMES 18
6. DIGITAL COLLECTIONS AT CORNELL UNIVERSITY LIBRARY 19

SECTION II. IMMEDIATELY ACTIONABLE STEPS 22

7. PRE-PRESERVATION ANALYSIS & PLANNING 22
8. IMPLEMENTING BACKUP AND BYTE-REPLICATION 25

SECTION III. PRESERVATION OF DIGITIZED BOOKS AND OTHER DIGITAL COLLECTIONS 27

9. DIGITAL PRESERVATION 27
10. REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) 30
11. A MODEL FOR CULTURAL HERITAGE ORGANIZATIONS 33
12. IMPLEMENTATION CHOICES 35

SECTION IV. APPENDICES 46

13. APPENDIX: GLOSSARY 46
14. APPENDIX: PARTICIPANTS IN THE PORTICO LOCALLY CREATED CONTENT STUDY 51
15. APPENDIX: PARTICIPANTS IN THE JISC PRESERVATION STUDY 52
16. APPENDIX: STRAW-MAN DESCRIPTION OF POSSIBLE PORTICO PRESERVATION SERVICE FOR LOCALLY CREATED CONTENT (LCC) 53
17. APPENDIX: TEMPLATE PRESERVATION POLICY 57
18. APPENDIX: ILLUSTRATIONS OF ANSWERS TO THE PRACTICAL QUESTIONS 58
19. APPENDIX: SOFTWARE SYSTEMS IN USE ACROSS BOTH STUDIES 71
20. APPENDIX: WORKSHEET TO ESTIMATE COSTS 72

Section I. INTRODUCTION AND RESEARCH

1. INTRODUCTION

Over the past decade there has been tremendous growth in the number of digitization projects initiated by cultural heritage organizations. These organizations have long been the stewards of our creative and scientific outputs and have been taking advantage of digital technologies to make their content broadly available.

In 2010, OCLC Research executed a survey of 169 institutions about their special collections as a follow-up to the survey of special collections executed by the Association of Research Libraries (ARL) in 1998 (one outcome of that earlier survey was a webpage on the ARL website highlighting the unique role of special collections[1].) The institutions surveyed by OCLC Research included members of the following organizations:

- Association of Research Libraries (ARL)
- Canadian Association of Research Libraries (CARL)
- Independent Research Libraries Association (IRLA)
- Oberlin Group
- RLG Partnership (U.S. and Canada)

Of those institutions surveyed by OCLC Research in 2010, "ninety-seven percent (97%) have completed one or more digitization projects and/or have an active program."[2]

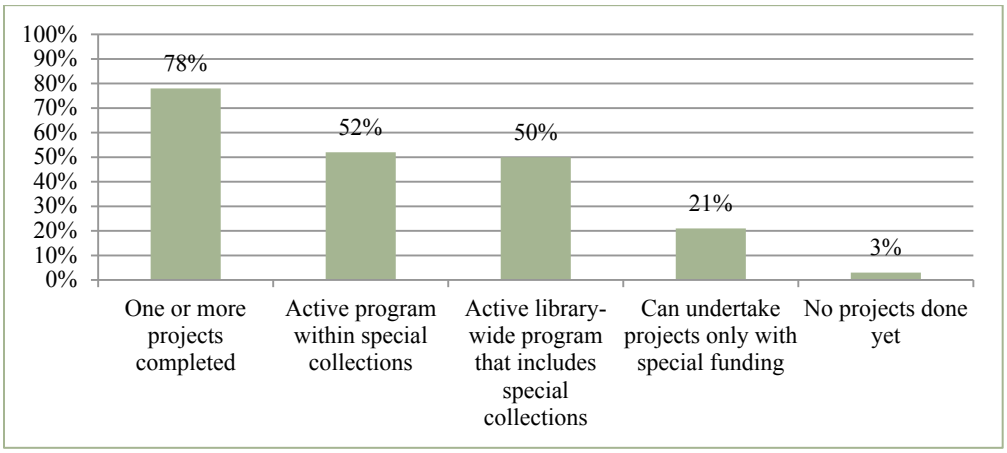


Figure 1: OCLC Survey Results – Digitization Activity (Dooley, 2010, p. 54)

Digitization has long been associated with preservation, as it is one tool an archivist can use to preserve fragile content. The Society for American Archivists teaches an entire course on

[1] http://www.arl.org/rtl/speccoll/

[2] Dooley, Jackie M. and Katherine Luce. "Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives." OCLC Research. Dublin, Ohio: 2010. Available at http://www.oclc.org/research/publications/library/2010/2010-11.pdf and last accessed on Mar 31, 2011.

"Digitization for Preservation"[3] and in 2004 the ARL Preservation of Research Library Materials Committee commissioned a report on "Recognizing Digitization as a Preservation Formatting Method."[4]) The widespread adoption of network technologies and the digitization of a diverse array of primary and secondary materials have opened to scholars, researchers and students unprecedented access to a vast array of materials vitas to advancement of the humanities. The broadened access that robust networks and digitization have enabled has strengthened enormously the prospects for continued robust advancement of the humanities which rely so heavily upon access to a diverse and rich array of materials from the past. This essential access is threatened as the corpus of digitized materials grows, and this significant new risk stems directly from the special vulnerability of digital objects. Unlike physical objects—books, letters, or manuscripts—which under reasonable conditions can last for many decades with only minimal attention, digital objects are extremely short lived unless intensive preservation attention is routinely provided. It is beginning to be understood that the substantial investment cultural heritage organizations are making in creating digital collections must be met with a commitment and infrastructure to protect this content for its lifetime. For example, JISC, an organization that inspires UK colleges and universities in the innovative use of digital technologies, now requires the digitization projects it funds to develop a preservation plan for the digitized content.

Portico Research

In one response to this need to develop models of digital preservation, the NEH and IMLS awarded a grant to Portico, in partnership with Cornell University Library, through the "Advancing Knowledge: The IMLS/NEH Digital Partnership grant program" to develop a practical model for how preservation can be accomplished for digitized books. Through this initiative and other efforts, Portico had the opportunity to discuss digital collections and their long-term preservation with 27 cultural heritage organizations. In addition, Cornell University Library provided significant samples of content to analyze. Out of this research and the extensive experience in preservation at both Portico and Cornell University Library, we developed a model for the preservation of digitized books and other "document like" digital content at cultural heritage organizations.

[3] http://www2.archivists.org/dae/university-of-michigan/digitization-for-preservation

[4] http://www.arl.org/bm~doc/digi_preserv.pdf

2. Are the steps we're taking to safeguard our content sufficient?

Cultural heritage organizations may find themselves contemplating their digital collections and wondering if their content is protected right now. This high-level consideration manifests itself through questions like the following:

- The IT department backs up the server, isn't that sufficient?
- We make a tape backup every 3 months, are we covered?
- The high resolution master files are on an external drive in Joe's office, is that OK?
- Can we keep this collection safe without preserving it?
- What will make this digital collection "safe enough"?

There are no single answers to the questions posed above. Rather, answers are dependent upon the needs of the collection, the content owners, the users, and the cultural heritage organization. However, these questions are an excellent place to begin considering short- and long-term digital preservation needs and options.

What are the Protection and Preservation Choices?

The methods a gardener uses to 'preserve' strawberries for the coming winter months are quite different from those a plant biologist uses to save specimens for study over the coming decades. So, too, the methods used to protect content for use in the near-term differ from those used to preserve content over the long-term.

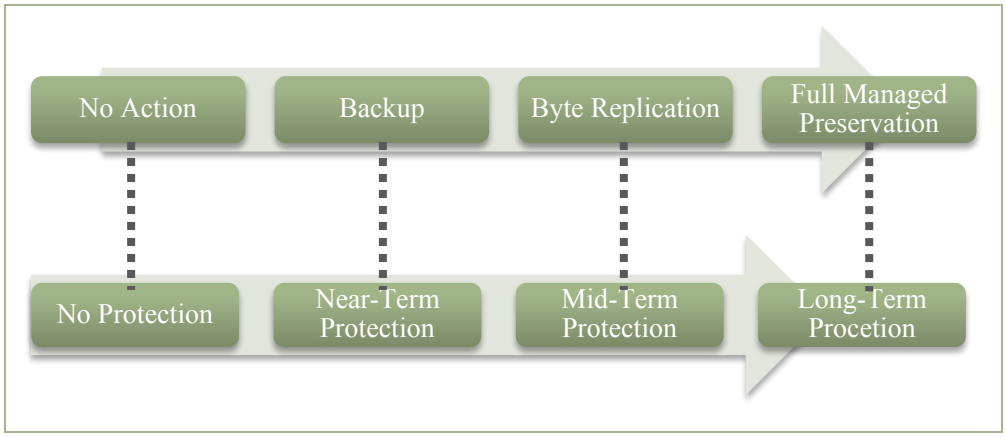


Figure 2: Protection and Preservation Continuum

The options for protecting access to digital content can be placed along a continuum (see *Figure 2: Protection and Preservation Continuum* above) that concludes in full digital preservation and long-term protection of access to content.

Backup provides near-term protection: Backup—when content is copied and stored in multiple locations to create readily available data replacements in case of equipment failure or other catastrophe—has long been understood to be a requirement for protection of near-term data access. It is imperative for business continuity and it is necessary to ensure that access to content in the near-term will not be interrupted for any length of time. A well-managed backup system can help quickly resolve problems with content encountered this week, or next week, or next month, but not over the long-term. Backup is typically implemented with commercial software that allows users to retrieve files backed up at specific points in time. Very often, content may only be retrieved via the software with which it was originally backed up. If special software or hardware is required to access the content and if it has been compressed via a proprietary technology, the long-term future accessibility and authenticity of the content—key goals of digital preservation—cannot be assured.

Byte Replication provides mid-term protection: Byte replication is a process whereby identical, multiple copies of files, file systems, or websites are created. The copies may be written to other online computers or to offline media. These replicas are typically held in diverse geographic locations and specialized software is not needed to access the content. This diversity in copies and location, together with the lack of reliance on software, ensures that byte replicas should provide content that is authentic and usable for as long as the file formats remain usable. However, simple byte replication includes no provision for ensuring the content is usable when the file formats are no longer current, nor is there any inherent provision for ensuring that the content remains discoverable. For example, if a series of book files are byte-replicated without accessible bibliographic information describing the intellectual content of the replica, there is no guarantee that an end user in the future will be able to find the specific article he or she needs.

Full Managed Digital Preservation: Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long-term. The key goals of digital preservation include:

- usability — the intellectual content of the item must remain usable via the delivery mechanism of current technology
- discoverability — the content must have logical bibliographic metadata so that the content can be found by end users through time
- authenticity — the provenance of the content must be proven and the content an authentic replica of the original as deposited
- accessibility — the content must be available for use to the appropriate community

In order to successfully perform full managed digital preservation as defined above, an organization must meet the following requirements:

- A mission to carry out preservation—as noted in the CLIR survey, *E-Journal Archiving Metes and Bounds: A Survey of the Landscape*, the first indicator of an archiving program's reliability is that it "have both an explicit mission and the necessary mandate to perform long-term … archiving."[5] The mission creates an environment conducive to the specialized planning and infrastructure needed to support digital preservation.
- A sustainable economic model to support the preservation activities over the identified lifetime of each digital collection.
- Clear legal rights to preserve the content.
- A relationship with the content provider or copyright owner, as it is often necessary to discuss the content and what preservation actions are appropriate to be taken on it with the copyright owner.
- Relationships with the users of the content, such that the cultural heritage organization can ensure it is meeting the needs of its users.
- A preservation strategy consistent with best practices and a technological infrastructure able to support the selected preservation strategy.
- Transparency about the organization's preservation services and strategies, clients, and content.

It is worth noting that backup and byte replication are required elements of long-term preservation and thus are appropriate first steps in protecting content for long-term access through preservation.

What is the Right Choice?

For an organization that is only beginning to contemplate and plan for long-term digital preservation, it is often best to take an incremental, step-wise approach. The most important initial measures include:

1. **Locate all the content:** It is common for content to be widely dispersed at cultural heritage organizations, with the master copy of the metadata located in one place, the high resolution master files in another place, and both separated from the derived copies of the metadata and content files (which typically live together in a repository system).

[5] Kenney, Anne R., Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern and Ellie L. Buckley. "E-Journal Archiving Metes and Bounds: A Survey of the Landscape." Council on Library and Information Resources: Washington, DC (2006). Available at http://www.clir.org/pubs/reports/pub138/pub138.pdf and last accessed on Mar 31, 2011.

2. **Initiate regular backups**: Once all the content has been identified and locations documented, organizations should ensure that regular backups are being made of the content.
3. **Test retrieval from backups**. A backup is only worthwhile if content can be retrieved from it.
4. **Develop a long-term preservation plan**: The questions provided in
5. *Pre-Preservation Analysis* & Planning below are intended to assist with this process.

When an organization is ready to begin long-term preservation, the form that such preservation takes will depend upon many factors, including the length of time for which the content must remain usable and the collaborative arrangements the organization may choose to make. For more on this topic see the section on *Preservation of Digitized Books and other Digital Collections.*

3. Research and Analysis

Between fall 2008 and fall 2009, Portico spoke with 27 cultural heritage organizations, including a range of educational institutions, national archives, national libraries, and museums, about their locally created content through two projects:

1. Portico Locally Created Content Study (LCC): Through this NEH-funded study, which is the focus of this paper, Portico worked with Cornell University Library and a group of librarians from 14 additional libraries at institutions of higher education to evaluate the technology and costs associated with preservation of locally created digital content (born digital and digitized) that is maintained by the institution. The librarians participated because they were investigating preservation solutions to apply to their digital collections. The diverse group of librarians included Portico participants and non-participants, representation from a number of countries, and librarians from schools of varying sizes and types. Using a template as a guide, Portico staff and the librarians individually discussed their content and needs. Out of these discussions, Portico developed a straw-man preservation service and discussed this model with the institutions (see *Appendix: Straw-man Description of Possible Portico Preservation Service for Locally Created Content (LCC)*). We also analyzed possible estimated costs associated with the service.

2. JISC Preservation Study: Through this JISC-funded study, Portico partnered with the Digital Preservation Coalition (DPC) and the University of London Computing Centre (ULCC) to carry out an extensive analysis of 16 projects funded through the JISC Digitisation Programme. The ULCC staff interviewed each project using a template to guide the discussion. Portico then reviewed the gathered data and the preservation plans for the 16 JISC digitisation projects. The results of this work were a private report to JISC describing the specific risks and recommendations for each project, a public report describing strategic risks and recommendations for JISC to address in future funding, and four detailed case studies.[6] We reference this research work in this paper because it contributed to the overall knowledge base Portico brought to the effort to develop a model of preservation of digitized books for cultural heritage organizations.

[6] http://www.dpconline.org/advocacy/knowledge-base/594-digitisation-programme-digital-preservation-study

4. Overview of Cultural Heritage Organizations and their Projects

Collections

The collections developed and maintained by the institutions were as varied as the institutions themselves, from the art museum implementing an early digitization project of its own collections to collections of digitized library special collections to massive digitization efforts at national libraries. The institutions are collecting and making available a large variety of content, including:

- Digitized special collections content and ephemera (these can include images, OCR text, audio/video, etc.), including:
  - Photographs (historical and art slide collections), glass plates, photographic negatives, maps, illustrations, postcards, posters, playbills, theatre programs, prints, and architectural images
  - Digitization of prints and drawings on paper; paintings on canvas; stained glass; costumes; letters; wood blocks; tapestries; and art objects
  - Collections about the history of the institution including president's reports, photographs, and senate minutes
  - Out-of-copyright books
  - Collections of letters, diaries, sheet music, medieval manuscripts, historical newsletters, rare books, scripts, letters, architects' plans, press cuttings, and pamphlets
  - Multilingual collections of texts
  - Manuscripts, reports, and state and local documents
  - Historic newspapers, current newspapers, historic and current journals
  - Glass plate negatives and photogravure plates
  - Brain scans and x-rays (under discussion at one institution)
  - A/V digitized from the institutional archives, video records of theatre, public record films, parliamentary coverage, national news broa*dc*asts, and campaigning films
  - Oral histories (recordings and transcriptions), field recordings, news recordings, and music recordings
- Digitized maps (some georeferenced to surveys), historical and current data about local governmental units, digitized gazetteers & digitized books and documents, geographically located historical statistics
- Institutional publications – current and historical
- Electronic Thesis and Dissertations (ETDs), senior thesis work, technical reports, working papers, faculty and student research and publications, and grey literature

- Class lectures in video and audio, recordings of talks and readings, educational videos, and webcasts of special events and campus-wide events
- Learning objects, web video, and flash and Camtasia tutorials
- Datasets

Most institutions were digitizing their existing physical special collections and ephemera and nearly all collections discussed were curated (even those collections that were traditional institutional repository type content). Figure 3 below shows the percentage of institutions working with each content type.

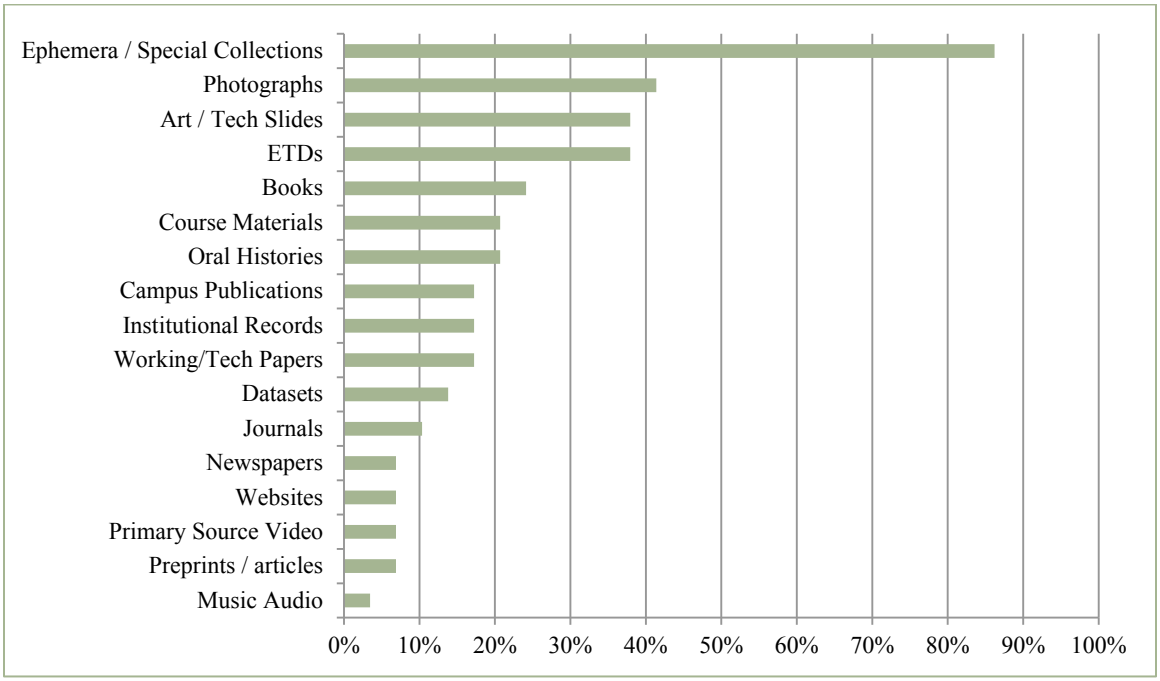


Figure 3: Percentage of organizations indicating that there are working with these types of materials

The materials collected by study participants roughly approximate the result of an ITHAKA survey of libraries from institutions of higher education in the United States that was performed in 2006.

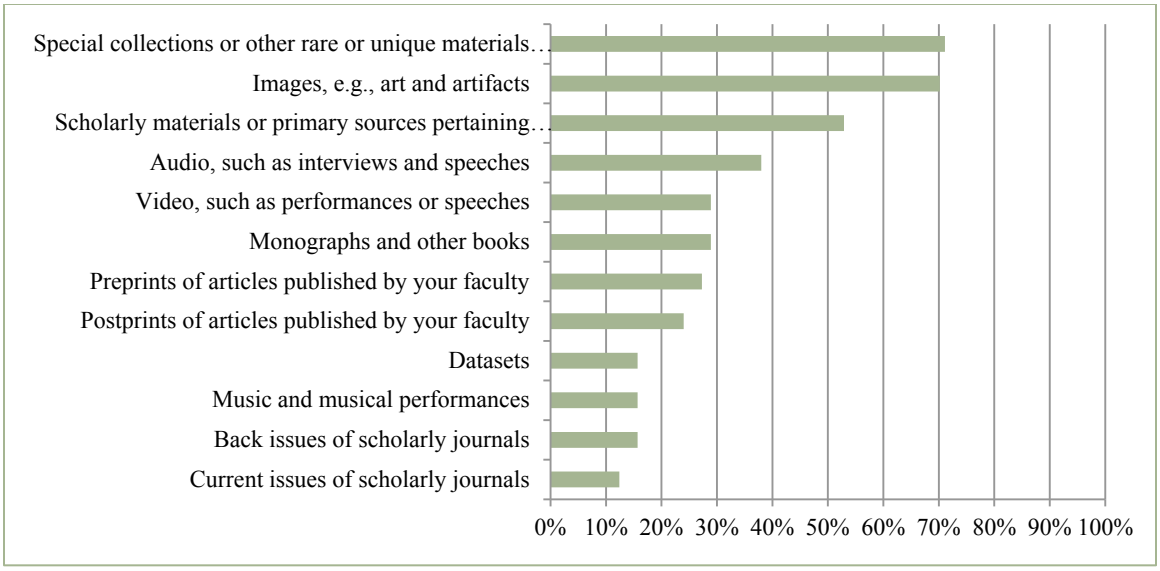


Figure 4: Percent of librarians indicating that their digital repositories contain these types of materials[7]

Most institutions have both open collections to which they intend to continue to add content over time, and closed collections to which no additional content will be added. Most institutions believe that the amount of digital content in their care will continue to grow—either through the addition of new collections or the addition of content to existing collections.

Content Management

Institutions are using a variety of content management tools. The 27 institutions, and at least 30 projects, reviewed were using:

- 36 distinct pieces of software
- 94 instances of software
- 3.13 pieces of content management software, on average

The systems used by the institutions reviewed included:

- Image Repository Systems
  - MDID
  - Luna
  - Artesia
- Third Party Delivery
  - JSTOR
  - Cengage
  - ProQuest

[7] Housewright, R., & Schonfeld, R. (2008). *Ithaka's 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education*: Ithaka Retrieved Dec 10, 2008, from http://www.ithaka.org/research/Ithakas%202006%20Studies%20of%20Key%20Stakeholders%20in%20the%20Digital%20Transformation%20in%20Higher%20Education.pdf

- Repository Systems
  - CONTENTdm (local, hosted, pro)
  - Fedora
  - DSpace
  - ExLibris DigiTool
  - Innovative's Symposia
  - BePress Digital Commons
  - VITAL
- Search Tools
  - Solr
  - DTSearch
- Audio/Visual Systems
  - iTunesU
  - Streaming Server
- File Server
- Journal Delivery System
- Catalog Systems
  - IRIS (MD in FMPro)
  - CALM
  - MODES Catalogue
  - Unknown Catalogue
  - Extensis Portfolio
  - Relational Databases
  - Allegro Database
  - OPACs
  - Tec-Rec
  - SIFT
  - MINISIS
- Preservation Systems
  - Bespoke Preservation
  - Quantum Digital Archive

Most institutions surveyed were using one or more repository systems. Smaller institutions were as likely to have digital content as larger institutions. Smaller institutions often did not have staffing to allow them to run a local repository like Fedora or DSpace and they were more likely to use a hosted service (predominantly, CONTENTdm, but also including consortia-based repositories). Figure 5 below charts the types of systems in use and how many instances of each were used across the 27 institutions we surveyed.

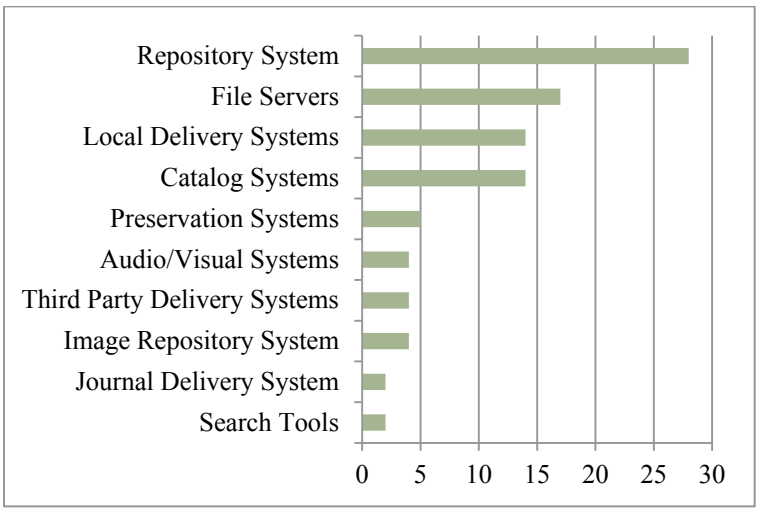


Figure 5: Number of Individual Instances of Types of Systems in Use across all Organizations Analyzed

Many institutions, even those using a hosted repository, have stacked software systems one on top of the other in order to coordinate the full spectrum of functionality they need to manage this content. For example, some typical "stacks" include:

- an external drive for high resolution master files
- a content management system for managing the metadata and some delivery objects

In addition, many institutions manage:

- an image server
- and/or a streaming server for delivery of specialized content
- a catalog system (perhaps already extant for more traditional physical resources)

A common thread running through the range of content and institutions is that the high resolution master files (high resolution images, high quality audio or video, etc.) are not collocated with the repository or delivery content management system. At a majority of institutions, these master copies are loosely coupled to their delivery objects and metadata through naming schemes, spreadsheets, or databases.

Workflow

The cultural heritage organizations that we spoke with all followed a similar set of steps to digitize and curate the content. In general, their digitization and curation processes were manual and the workflow was managed through a spreadsheet or other checklist.

For most institutions, the content management process—managing how content moves from one location to another during the digitization and deposit process and then managing the ongoing maintenance of content—is the most difficult and time consuming aspect of the project. Even institutions that have a repository or a content management system typically store their high resolution master files on a file server. Files and metadata that are not collocated inevitably have a tenuous connection and are at much greater risk to become unsynchronized. Organizations with the master metadata and high resolution master files collocated are in a much stronger position to begin digital preservation.

Very few institutions that we spoke with can package content files together with their matching metadata files and move that content from place-to-place. The standard repository systems, today, do not provide this service as OAI-PMH is not sufficient for data transfer. This inability is another risk factor to the long-term preservation of content, as without this functionality and ability the content is locked into existing systems and likely to become unsynchronized.

Collaboration

All institutions in our discussions were collaborating with other departments within their own organization or with outside organizations. This collaboration has many manifestations from outsourcing the digitization to providing a collaborative delivery service.

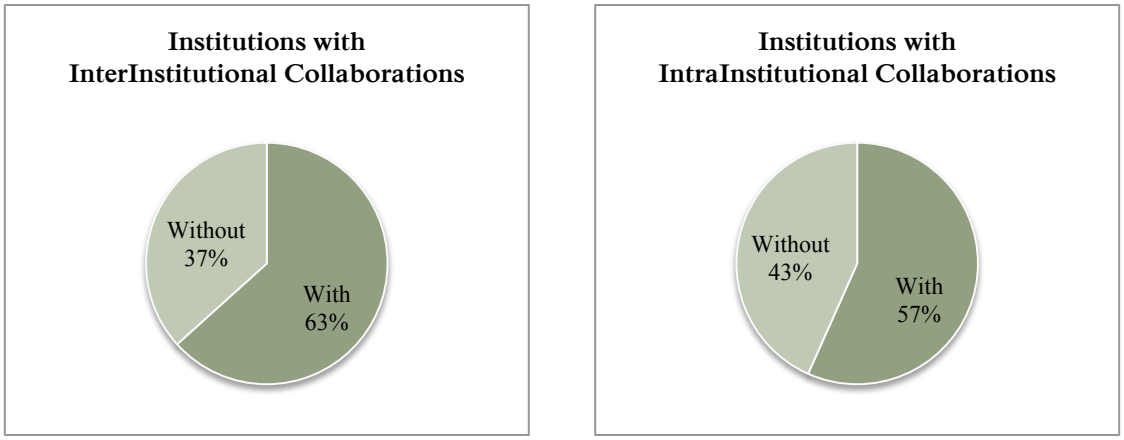


Figure 6: Charts of Inter and IntraInstitutional Collaboration

The survey results published by the Primary Research Group in their 2011 Survey of Library & Museum Digitization Projects found similar results, where "more than 54% of survey respondents have teamed up with some other department of their institution to work jointly on a digitization project" and "51% of the institutions sampled have outsourced to a third party some aspect of their digitization efforts."[8]

Collection Management and Content Management

There are two types of ongoing management needed to successfully maintain collections of digitized or otherwise digital content.

Collection Management:

- Collection management is the set of activities necessary to maintain the intelluctual units being preserved. This includes initial curation of and creation of descriptive metadata for the content, initial scanning or other digitization, and management of the creation workflow. Collection Management also includes ongoing activties that will continue as long as the collection remains availbale, such as correcting errons in the preserved content.
- Many digitization projects are planned to be static and are funded with one-time money allocated to digitize the content and make it available for use. Despite best intentions, however, most collections are not static and, over time, collection management continues to occur as content must be added, updated, and deleted.

[8] Primary Research Group. "Survey of Library & Museum Digitization Projects – 2011 Edition." (2010). Available at http://www.primaryresearch.com/view_product.php?report_id=281 and last accessed on Mar 31, 2011. Pages 33 and 34.

Content Management

- Content management includes those activities targeted at managing the files and databases that comprise the intellectual units being preserved, for example: backups, ongoing fixity checks to ensure there has been no degradation to the files, media migration when the old media on which content is stored needs to be replaced, file migration if the formats of the files become unusable, and other activities.

These two activities should be considered separately, yet we found that many institutions interviewed often conflated or ignored these activities. Ongoing collection management requires the skills of a subject specialist to determine when and how to update and curate the content files and metadata, whereas ongoing content management requires technical skills to replace aging hardware and migrate files from old formats to new formats.

5. THEMES

A number of themes arose from the analysis of the interviewed institutions and their content:

The size of a cultural heritage organization is neither a predictor for risk nor amount or quality of digital content.

Most organizations work with external or internal partners for digitization, delivery, and/or preservation. Very few are working entirely independently.

Few cultural heritage organizations have easy access to their high-resolution master files, as such files are typically on DVD, CD, or external hard drive and not in the repository. It is preservation of these files that is most important, as they are the items most expensive to reproduce.

Externally held files, such as the high resolution master files, often have a very tenuous connection to their metadata. Without a tight coupling to metadata, the files will be unusable in the future.

Most of the cultural heritage organizations with which we spoke cannot package up the high resolution master files, the derived files, and the metadata and move the package as a unit from one system to another as required to meet the definition of full, managed preservation.

Most cultural heritage organizations do not have strong digital content management processes and control (e.g., nearly all organizations surveyed used a file server as one element of their content management strategy, in addition to other content management systems) and this puts their content at risk.

Many cultural heritage organizations do not have staff to support either preservation or access systems in-house.

The repository systems in place today cannot package and transfer content in a standard format. Most have OAI-PMH functionality, but OAI-PMH is not sufficient for data transfer.

Analysis shows that many cultural heritage organizations would benefit from a turn-key solution that provides both access and preservation for a large variety of formats and content types. Such a "one stop shop" would be cost-effective for institutions with a need for protection, but less rigorous preservation and access.

Cultural heritage organizations are acting as publishers (for example, in 2001 the University of Michigan Library opened the Scholarly Publishing Office), but this is not a traditional business for them.

Cultural heritage organizations do not often have a sustainability plan associated with their digital content. Rather than considering the digital content to be a product that must be sustained, it is considered another outlet for their special collections. It is not clear if their parent institutions will think of this content in the same way.

6. Digital Collections at Cornell University Library

The participation of Cornell University Library on the NEH/IMLS grant allowed Portico and Cornell University Library to perform an in depth analysis of the Library content in the context of digital preservation. Cornell University Library has been creating and managing digital content since the mid-1990s, including numerous digitization projects managed in house and externally. In addition, the Library authored the highly valuable Digital Imaging Tutorial,[9] which has been used by many diverse organizations as they implement their own digitization and digital preservation endeavors. The Library was also the original author of Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions.[10] Cornell University Library provided in depth advice on analysis of the Portico locally created content survey and development of the straw-man preservation model that Portico shared with the institutions for evaluation as part of its work to develop a model suitable for cultural heritage organizations.

Cornell University Library provided samples of digitized content to Portico for analysis, including samples of early digitization projects, including:

- Ezra Cornell Papers[11] consisting of TIFF images, GIF images, OCR and coordinate information, a manifest file, and descriptive metadata encoded in RFC 1807[12]
- Making of America[13] consisting of TIFF images, GIF images, OCR and coordinate information, and descriptive metadata encoded in both RFC 1807 and in the EFFECT Technical Specifications[14] file which also contains manifest and structure information
- Southeast Asia Visions[15] consisting of TIFF images

[9] http://www.library.cornell.edu/preservation/tutorial/contents.html

[10] http://www.icpsr.umich.edu/dpm/

[11] http://historical.library.cornell.edu/ezra/browse.html

[12] RFC 1807 is an IETF request for comment memo issued in 1995 that describes a formant for describing technical records (see http://tools.ietf.org/html/rfc1807).

[13] http://ebooks.library.cornell.edu/m/moa/

[14] The "EFFECT" Exchange Format For Electronic Components and Texts Technical Specifications was developed by Elesvier in conjunction with a number of early digitization projects at Universities, including Cornell and the University of Michigan. It provided a format to encode descriptive metadata and packaging information about hierarchical, serial publications (such as scholarly journals). A copy of the EFFECT specification is still available at http://www.info.sciverse.com/UserFiles/Files/sciencedirect/effect40.pdf (as of March 2011).

[15] http://digital.library.cornell.edu/s/sea/index.php

Samples of middle year digitization projects, including:

- Historical Math Monographs[16] — consisting of TIFF images, OCR and coordinate information, descriptive metadata in XML, and a manifest file
- Samuel J. May Anti-Slavery Collection[17] — consisting of TIFF images, GIF images, OCR files, metadata and full-text in SGML, a manifest file, and metadata in XML

And samples of more current digitization projects, including:

- The Cornell Daily Sun[18] — consisting of PDF files, TIFF image files, METS XML files containing descriptive metadata for the issues and articles, issue structure information, and file manifests, and XML files containing OCR and coordinate information
- Microsoft Digitization scanned books[19] — consisting of JPEG2000 files, a MARC metadata file in MARC, MODS, and MARC XML, manifest files, OCR files, and an XML Dublin Core metadata file

This content comprised a fascinating swath of digitization experiences, as it was created over the course of nearly two decades. The Cornell content provided an excellent exemplar of the varieties of content, file formats, and metadata formats that exist at cultural heritage organizations. Portico looked at content from three Cornell projects in depth to analyze their preservation options within the context of the Portico archive: books digitized by Microsoft, the Papers of Ezra Cornell (a very early digitization project at Cornell), and Cornell Daily Sun.

Per the straw-man service model (see *Appendix: Straw-man Description of Possible Portico Preservation Service for Locally Created Content (LCC)*), Portico considered two options for each Cornell collection:

Zip and Hold:

- Package content into ZIP files and hold it in the archive
- Perform standard archive maintenance of on- and off-line replication, on- and off-line media refreshment, fixty and completeness checks, receipt and processing reports, audit accreditation reports, and regular status reports on holdings, repairs, fixity, completeness and migrations

[16] http://digital.library.cornell.edu/m/math/index.php

[17] http://digital.library.cornell.edu/m/mayantislavery/

[18] http://cdsun.library.cornell.edu/cgi-bin/newscornell

[19] http://www.archive.org/details/cornell

Full Preservation Activities:

- Analyze the structure of the content to determine whether all expected files were received
- Validate files against their format specifications and revalidate files in the future as new tools are developed
- Repackage content into an archival information package (AIP)
- Migrate files to new formats on ingest or in the future as necessitated by the changing technological environment
- Perform standard archive maintenance of on- and off-line replication, on- and off-line media refreshment, fixty and completeness checks, receipt and processing reports, audit accreditation reports, and regular status reports on holdings, repairs, fixity, completeness and migrations

Because of the quantity of content in the Microsoft digitized book project and concerns about the consistency of the data, Portico recommended protection of the Microsoft digitized book project via Zip and Hold, whereby key Dublin Core metadata descriptors would be culled from the provided descriptive metadata files and placed in the preservation metadata files without further validation of the metadata files. In addition, we recommended Zip and Hold for the Papers of Ezra Cornell. The Papers is an early digitization project and while the images and structure of the packaging are very clean, the metadata is in a format which is difficult to validate. The Cornell Daily Sun, however, is one of the later digitization projects from the Cornell University Library and in addition to beautiful TIFF files, it has well-constructed XML metadata files that can be validated and thus for this collection Portico recommended Full Preservation Activities. Portico processed a number of issues from the Cornell Daily Sun digitization project and was able to successfully validate the XML files and repackage these entire issues into an archival information package suitable for preservation within the Portico archive. The development of this tool set and processing of the content took approximately one month of one developer's time.

One of the key lessons learned by both Portico and Cornell in regard to the preservation of the library content is that to implement long-term digital preservation via the Full Preservation model (rather than the back-up and byte replication of the Zip and Hold model) would require the development of specific tools for each Cornell collection. This matches Portico experiences with e-journals, e-books, and d-collections where, despite the presence of standards within these communities, each "collection" requires a tailored suite of tools. This also conforms to our understanding of the varied content at other cultural heritage organizations and suggests that in order to be cost effective, the protection of this content may need to be managed through the means of less customized tools.

Section II. Immediately Actionable Steps

7. Pre-Preservation Analysis & Planning

As a result of this analysis, institutions will be able to make informed decisions about the length of time the collection must be protected and therefore the amount of investment to be made in that protection (for example, is backup and/or byte replication sufficient, or does the collection need long-term, managed digital preservation.) The report created through answering the following questions can also be used to form the basis of a preservation policy for the content.

Who: Identify the key players involved with long-term preservation of the targeted content	
Our surveys have shown that, especially when multiple partners are involved in managing content, there is an opportunity for misunderstandings to arise around which party is responsible for which element of the content. Often the role responsible for managing the files is different from the role responsible for managing the intellectual content of the collection. Therefore, for each digital collection, it is important to identify the key players involved with the development and long-term management of the content.	1. Who is writing the policy and plan? 2. Who has responsibility for maintaining the intellectual content of this collection (e.g. making corrections to metadata or content files)? Who has curation responsibilities and is the advocate for the collection? 3. Who has responsibility for maintaining the bytes of the files in this collection (e.g. identifying and fixing corrupted files)? 4. Who approved this policy and plan? 5. Who will use the content in the short and long-term?

What: Describe or characterize the collection and content	
Per the definition of digital preservation, being able to trace the authenticity of an object in the collection is important. From a practical point-of-view, this provides information to those people who will be managing the content in the future, but may not have been involved in its original creation. The ability to quickly characterize a collection is also very important when it becomes necessary to consider all digital collections at one organization and organization-wide preservation solutions.	6. What is the content and from where did the content originate? 7. What file formats, including metadata formats, are present? 8. How many items are in the collection? How large is the collection on disk?

Where: Document the locations of all the copies of the content and metadata.	
Our surveys have found that there are often many versions of content "around." In order to manage all this content in a sensible way, it is important to identify where all the content is and the purpose of the copy at each location.	9. Where are the high resolution master copies of the descriptive metadata kept? 10. Where are the master copies of the content files kept? 11. Where are all the copies of the content, including backups, and how are the copies of the content related?

When: Document the targeted preservation timeframe and impact of loss.	
Not all content must be preserved forever, some content can be protected for a limited time, after which its status will be re-evaluated. Identifying what might happen if the content were irretrievably lost will help answer the question of how long it must remain available. Other factors include user demand and organizational mission.	12. How long should the content be available for use? 13. If the content is irretrievably lost, what are the repercussions?

How: Document how the key content management and preservation tasks will occur.	
It is important to make thoughtful decisions about how to manage the collection. A closed collection may be deposited into a read-only archive, whereas an open collection that will have updates made to it must be preserved in an archive that allows updates. Having all parties responsible for the content answer this set of questions together will ensure that everyone agrees on how the content will be managed.	14. How will the collection be created (perhaps draw a diagram of the workflow)? 15. How will the collection be maintained (perhaps draw a diagram of the workflow)? 16. Do you expect the content files to be migrated in the future? 17. May the content files be deleted? Added to? Updated? 18. May the descriptive metadata be deleted? Added to? Updated? 19. How will you track who did what and when to the content, if this is important to your organization? 20. How do you associate the master copy of the descriptive metadata with the high resolution copy of the content files and how will you move these two items around together?

See *Appendix: Illustrations of Answers to the Practical Questions* for example answers to the above questions. Additional self-assessment tools include:

- AIDA (Assessing Institutional Digital Assets) at http://aida.jiscinvolve.org/wp/.
- Drambora (Digital Repository Audit Method Based on Risk Assessment) at http://www.repositoryaudit.eu/ .
- TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist) at http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

8. Implementing Backup and Byte-Replication

After analyzing the content, the needs of the end users, and the resource base of the parent organization, some cultural heritage organizations may determine that a short or mid-term protection solution is sufficient for their purposes and may choose to implement that protection through backup and/or byte replication. In addition, those organizations which have determined they need longer-term protection, may choose to implement backup and/or byte replication while they are collecting and organizing their content in such a way as to make it possible to preserve it. Backup and byte replication will be elements of any long-term preservation solution and therefore taking these initial steps will build needed experience.

Backup and byte replication are well-understood solutions. Many cultural heritage organizations may be able to get robust backup from their parent institution. In general, backups provide solutions to two problems:

1. **User error recovery** – a user or system accidentally deletes or modifies some files and those few files need to be copied out of the backup and back onto the system. In this regard, currency of content is very important. The backup must have current versions of the files, or it cannot serve this purpose well. In order to support this type of file-by-file retrieval of current files and to quickly make the backups, most backup solutions implement a type of delta backup, such that only items that have been changed since the last backup are copied. The organization should expect retrieval of those few files to be relatively fast for the backup to effectively meet this need.
2. **Disaster recovery** – a natural or man-made disaster destroys the original copies of the content and the system needs to be entirely rebuilt. In this regard, currency is less important, as the organization will be spending considerable time rebuilding the system (perhaps even the machine room) and loss of a week or two of updates to the content will not significantly impact time to recovery.

Several things to consider when selecting a backup solution are:

The disks and software

- A frequent cause of data loss is failure of the hardware or media on which content is stored. This corrupts bits of the data. Enterprise disks often protect against this kind of loss by automatic detection and repair through the use of RAID and other software such as ZFS file system, but more commodity disks may not have these protection systems. In such scenarios, organizations may want to consider multiple backup solutions.

The location of the backup

- The farther away a backup is physically located from the original, the more secure it will be in case of natural or man-made disaster. However, often backups are used for day-to-day continuity management and speed of retrieval is important, in which case having the backups local is important.

Speed and frequency of backup	•Organizations should consider the amount of content to be backed up, the amount of time it takes to complete a backup, and the required frequency of backups. It cannot take longer to make a backup then the time allowed between backups.
Speed of retrieval	•When an organization needs to retrieve content from the backup, that speed of retrieval must match the needs the organization.
Monitoring	•An organization must identify who will be monitoring the backup and ensure that the backup system can provide reports in a manner that will be understood by the person responsible for the monitoring.
Testing	•Plan and execute regular test retrievals from the backup to ensure that the backup system is working as billed.
Technical skills of people responsible for the backup	•There are many backup solutions available today, from software built into external hard drives to on-line backup solutions in the cloud. Some of these solutions are plug-and-play and others require technical skills to implement. For example, a backup solution in the cloud may be implemented by licensing a service that does the backup automatically or by purchasing space from a service like Amazon S3 and writing your own backup software (the latter requires technical skills).
Proprietary formats	•Consider whether or not the backup solution rewrites content or simply copies the content byte-for-byte. There is a greater risk in using software that rewrites the content, rather than software that makes an exact replica of the content. However, an organization could decrease the size of their backups by using software that compresses the content.

Cultural heritage organizations should consider the benefits of multiple backups to address these tensions, including the following options:

- **Cloud Backup**: There are many on-line, cloud backup services available. This type of backup requires a high speed network and may be less reliable than other options due to transmission errors, difficulty in performing fixity checks, and less than 100% recovery guarantees from the backup service.
- **Off-Line Backup**: Creating backups to tape or external hard drives and shipping them to a secure, climate controlled environment, off-site environment is very reliable. The speed of content retrieval is slow, which makes it difficult for this type of backup to meet day-to-day business continuity needs.
- **Local Backup**: Backing content up to a local disk (full backups at regular intervals and incremental backups in between) is a third option. The speed of retrieval is fast, but reliability in case of huge, disaster is of concern, as the backup is located in the same general physical location as the content.

Section III. Preservation of Digitized Books and other Digital Collections

Cultural heritage organizations which are ready to begin the process of full preservation of their digitized books and other digital content need a model to follow, a place from which to begin the process. Included here is such a model that has been developed based upon our surveys, discussions with cultural heritage organizations, and our extensive experiences with digital preservation.

9. Digital Preservation

In order to support full digital preservation, an organization must devote time and attention to both the ongoing content and collection management of the preserved content. The preservation system must be monitored daily to identify system problems, the collection must be updated when errors in bibliographic metadata are found or when other problems with the intellectual content are identified, the internal and community understanding of file formats must be monitored, migrations of files to new formats must be performed, emulation software must be tested and preserved itself, hardware must be refreshed, and many other ongoing maintenance activities in order to support digital preservation:

Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. The key goals of digital preservation include:

usability	authenticity	discoverability	accessibility
• the intellectual content of the item must remain usable via the delivery mechanism of current technology	• the provenance of the content must be proven and the content an authentic replica of the original as deposited	• the content must have logical bibliographic metadata so that the content can be found by end users through time	• the content must be available for use to the appropriate community

Figure 7: Digital Preservation Definition

The four columns seen in Figure 7 above support long-term digital preservation and require investment by the cultural heritage organization.

Usability: Everyone has had the experience … that thesis written as a graduate student will no longer open. Maybe it is because it was stored on a floppy disk and maybe it was in WordStar. Maybe it was in WordPerfect and opens today, but the formatting is inaccurate. "Software designed for an older operating system may not run [on] its contemporary counterpart, which in turn means that files created using the software native to these older systems might now be

accessible on current computers. For example, a word processing document created in Windows 3.1 or Mac System 7.5 might not open with a modern office suite installed on Windows 7 or OSX."[20] File formats will become obsolete—it may take a long time and it may be that the files simply become more mangled in display than completely unusable, but it will happen.

Migration and emulation are the two primary strategies used for ensuring usability in long-term preservation. Migration involves transforming digital content from its existing format to a different format that is usable and accessible on the technology in current use. Emulation involves developing software that imitates earlier hardware and software. Migration is a strategy that requires a deep understanding of the content being preserved, whereas emulation is a more technology-based strategy, requiring a deep understanding of existing hardware and software. Within preservation policies, an organization should explain what preservation strategies are used for what content.

Cultural heritage organizations have a number of ways to address usability concerns, including:

For a succinct listing of file formats recommended for digital preservation and an explanation of why they are appropriate, see a handout created by the Florida Digital Archive[21].

Authenticity: Organizations engaged in digital preservation must prove that the current preserved objects are true to the item as originally deposited. Changes will be made to preserved content: descriptive metadata will be updated, files will be migrated, corrupted files will be replaced, etc. The cultural heritage organization or its preservation agency must closely track any changes made to the original preserved content in order to be able to continue to prove the current version is authentic to the original version. There are a variety of ways that this need can be met, including tracking changes through event records within the preservation metadata of the object or even keeping all versions of the content within the archive.

[20] Kirschenbaum, Mathew G., Richard Ovenden, and Gabriela Redwine. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections. Council on Library and Information Resources: Washington, DC (2010). p. 18. Available at http://www.clir.org/pubs/abstract/pub149abst.html and last accessed on Mar 31, 2011.

[21] http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Discoverability: In order to ensure the long-term digital preservation of any object there must be sufficient descriptive metadata associated with it to find it again. Within an archive, descriptive metadata is typically found in two places:

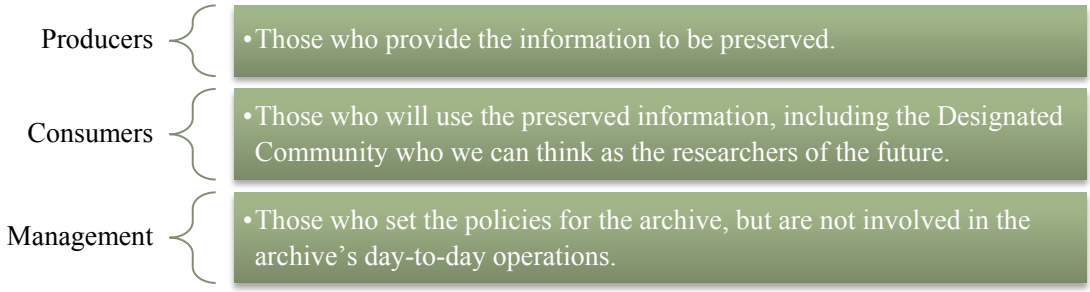
1. Encoded within the files that are the building blocks of the intellectual unit being preserved. For example, a digitized book may include an XML file that contains significant bibliographic information along with the full-text of the book.
2. Encoded within the archival system or preservation metadata files that provide a "wrapper" to the intellectual unit. For example, the record for that same digitized book in the archival system will have a minimal amount of descriptive metadata directly associated with it (so that archival administrative queries do not need to be made against the more complex and sophisticated XML file).

Accessibility: It is not enough for the cultural heritage organization or its preservation agency to keep the content safe and secure, they must be able to deliver that preserved content to users. Delivery requires a web service, uptime and response time requirements that may be different than those of the archive, user friendly search and browse functionality, and user support.

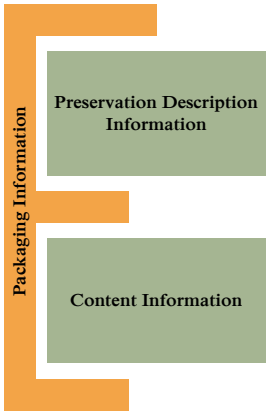
10. Reference Model for an Open Archival Information System (OAIS)

This recommended approach to preservation draws heavily upon concepts expressed in the Open Archival Information System (OAIS) framework[22]—the classic model that defines an archive as "consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community".[23] The biggest benefit from OAIS is that it provides parties with disparate backgrounds and concerns a shared terminology and as such, it is helpful for us to review the key assumptions and most useful constructs for cultural heritage organizations.

OAIS defines three categories of parties who have a vested interest in archival decisions, but who do not participate in the day-to-day management of the archive:



At the core of OAIS is the concept that each item preserved is an "Information Package" containing content information, preservation description information, and packaging information.



- **Content Information**
  - the original files to be preserved
  - information about the files being preserved so that they can be delivered to end users
- **Preservation Description Information** (the preservation metadata about the content information)
  - Provenance — where the content information originated, its custody chain, and event history
  - Context — how the content information relates to other information, outside itself
  - Reference — one or more identifiers, or systems of identifiers, by which the content information may be identified
  - Fixity — a checksum for the files
- **Packaging Information** (the metadata that logically binds all the elements of the content package)

Figure 8: OAIS Information Package

[22] http://public.ccsds.org/publications/archive/650x0b1.pdf

[23] Pg. 1-1

An information package is the unit of preservation. Conceptually, a single digitized book which is to become the information package might be built from the following files:

Figure 9: Files of a Digitized Book

The example digitized book above consists of scanned image files of the pages, image files of the figure graphics, an XML manifest file detailing how the page images should be sequenced together to create the book, a PDF version of the book, and a MARC record that encodes the bibliographic metadata about the book. All of these files and additional information on how the files all relate to one another must be preserved as the entire information package. Otherwise, all an organization has is image files on a disk or a MARC record without content.

OAIS assumes that there are three types of information packages:

Submission Information Package (SIP)	• The content that is delivered to the archive for preservation.
Archival Information Package (AIP)	• Through processing the submission information package is turned into an archival information package.
Dissemination Information Package (DIP)	• Through processing, the archival information package is delivered to end users or external system.

With these constructs in mind, we can consider the conceptual OAIS framework:

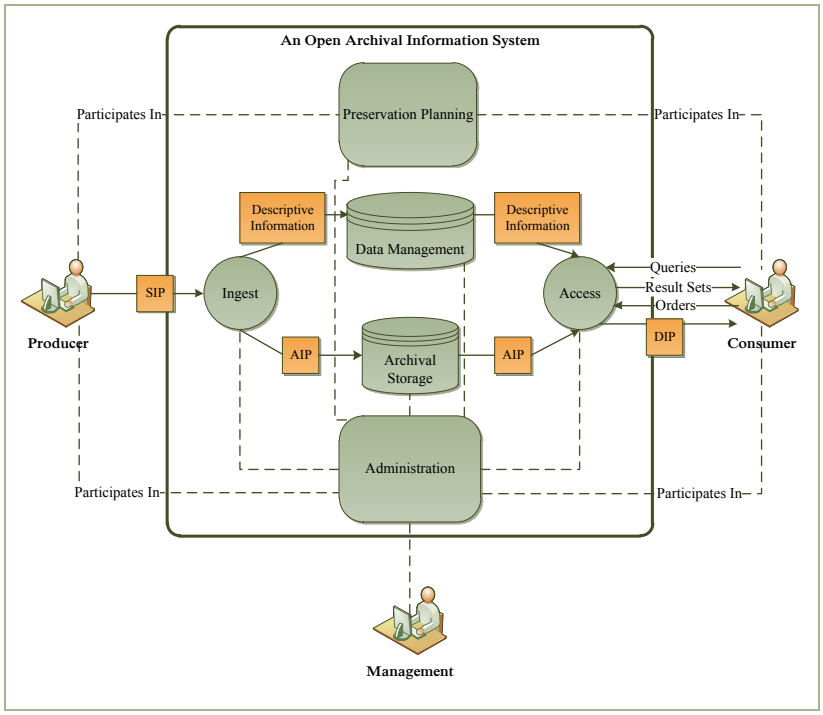


Figure 10: OAIS Functional Entities (CCSDS, 2002, pp. 4-1)

The most important ideas encompassed by the OAIS framework are:

- there must be preservation planning
- the archive requires ongoing administration
- content in the form of files and metadata about the files must come into the archive
- content in the form of files and metadata about the files must go out of the archive
- the archive is not the hardware or software—rather the hardware and software are elements of the archive, which includes activities performed by people
- the original producer of the content and the eventual users of the content must have input into the ongoing management of the archive
- at its very base, the preserved information consists of the content to be archived and metadata about the content—both of which are required for long-term preservation.

While OAIS is presented as one system, a number of preservation entities including California Digital Library (CDL) and PLANETS have implemented distributed OAIS compliant preservation services. CDL uses micro-services where curation and preservation functions are devolved "into a set of independent, but interoperable, services that embody curation values and strategies"[24] and PLANETS is implemented as a distributed service network.[25]

[24] "Curation Micro-Services" at the California Digital Library. Available at http://www.cdlib.org/services/uc3/curation/ and last accessed on Mar 31, 2011.

11. A Model for Cultural Heritage Organizations

The model proposed in this paper fits within the OAIS framework. In terms of the day-to-day actions of preservation, the following six activities are the key elements of the model:

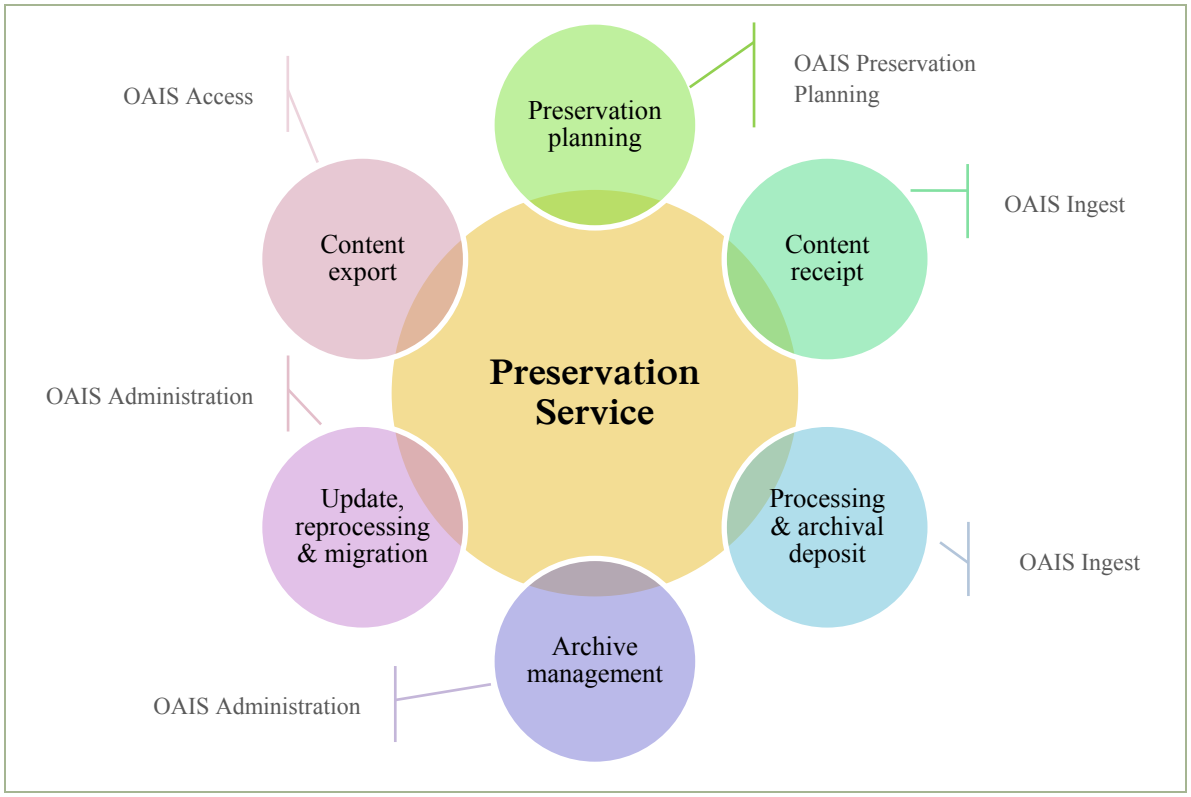


Figure 10: Preservation Model

Preservation Planning: The questions in

Pre-Preservation Analysis & Planning are a place to start when planning the preservation of a collection. As organizations move beyond backup and byte-replication they will need to develop formal preservation policies and sustainability plans (see *Implementation Choices* below.)

Content Receipt: Managing content before it enters the archive or repository is a key element of the six part preservation service model. Oftentimes cultural heritage organizations will manage the initial creation of content or receipt of content from others without the aid of a machine enforced workflow. Instead the workflow is managed by staff members or consultants using spreadsheets and checklists. While a machine enforced workflow is not a necessity, the process that moves content to the point of processing it into the archive must be tightly controlled.

[25] http://www.planets-project.eu/about/

Processing and Archival Deposit: It is unlikely that the content as originally received or produced is in the form of an archival information package. The steps involved in transforming the original content into an archival information package should be detailed in one of the preservation policies (either for the archive as a whole or a specific collection preservation policy, see *Preservation Policies and Review Process* for more information).

Archive Management: Ongoing monitoring and management of the archive must occur and includes a range of activities, including making and testing backups and byte-replication, monitoring for corruption and data loss, and monitoring the usability of the formats of files preserved in the archive.

Update, Reprocessing & Migration: Preserved content will be updated, reprocessed and migrated to meet both collection and content management needs. The preservation policies, organizational budget, and repository system must accommodate these ongoing processes.

Content Export: Cultural heritage organizations understand they must deliver content to end users, but it is also very important that the archive be able to package information packages up as units and move them out of the system en masse. This is the ultimate measure of the quality of an organization's content management.

12. Implementation Choices

Organizational Structure

Given the general six stage model above, one of the most important decisions is the organizational structure within which the content should be preserved.

Do-It-Yourself: In some instances, a cultural heritage organization may be able to provide its own digital preservation. The infrastructure required to provide adequate preservation is substantial, and thus this is an approach best taken by very large organizations such as national libraries, national archives, and institutions of higher education. For example, the British Library and the U.S. Government Printing Office both have developed substantial, internal preservation infrastructure.

Collaboration: Another option is collaborating with peer organizations to develop and maintain the processes and systems necessary for long-term preservation. This was a prominent approach across the projects we reviewed in our studies. Some examples of collaborative partnerships are:

- State wide collaborative repositories such as the GALILEO Knowledge Repository[26] in Georgia, being developed by Georgia Tech, the University of Georgia, Georgia State University, the Medical College of Georgia, Georgia Southern University, Valdosta State University, Albany State University, North Georgia College and State University, and the College of Coastal Georgia.
- Collaboration among peer institutions, such as the MetaArchive Cooperative[27] which currently has nearly 20 partners including such diverse institutions as Auburn University, Consorci de Biblioteques Universitaries de Catalunya, Florida State University, Folger Shakespeare Library, Georgia Tech, Historically Black Colleges and Universities Library Alliance, Library of Congress, and others.
- System wide collaborations such as the Digital Preservation Repository[28] of the California Digital Library (CDL) of the University of California.

Collaborative preservation systems can be built on a variety of platforms. GALILEO is built upon DSpace and MetaArchive upon a private LOCKSS network.

Third Party Preservation Service: Some cultural heritage organizations may choose to outsource their digital preservation to a third party digital preservation service such as Portico, HathiTrust, or JSTOR.

---

[26] http://www.library.gatech.edu/gkr/

[27] http://www.metaarchive.org/

[28] http://www.cdlib.org/services/uc3/dpr.html

Content Management & Technology

Although developing a preservation service is not, per se, a technology project, it is a content management project that relies heavily upon technology. We recommend a high-level workflow and configuration of systems as illustrated in Figure 9 below:

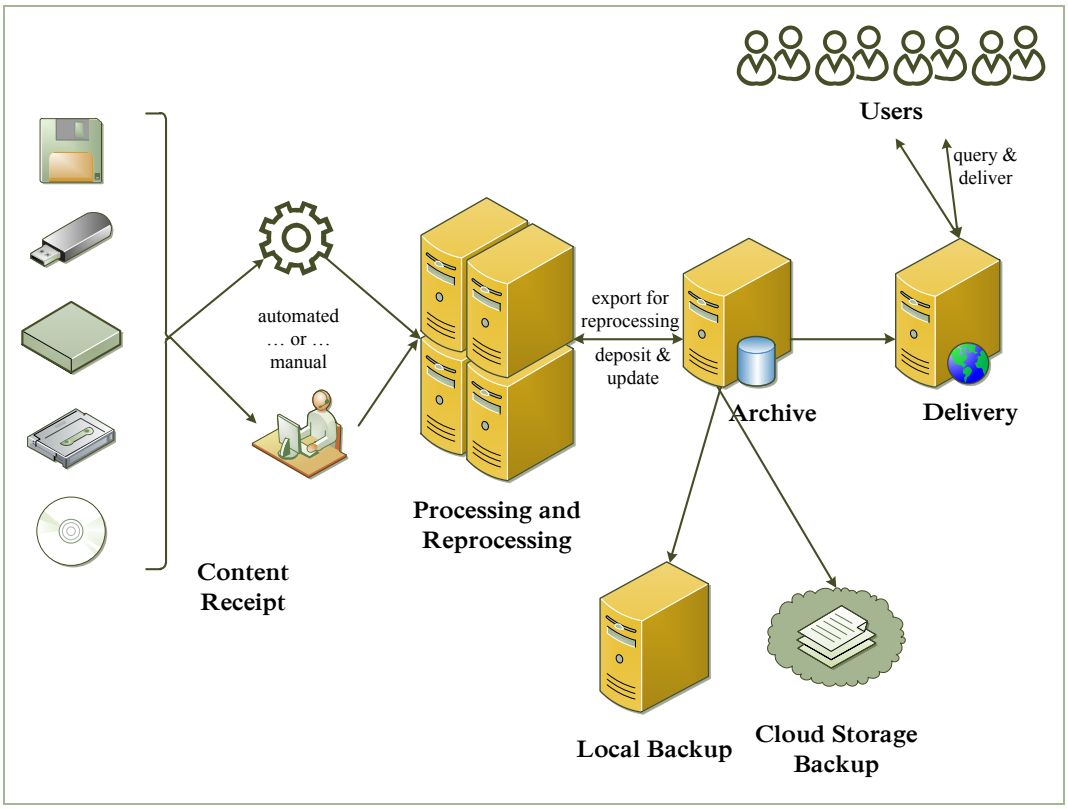


Figure 9: Recommended Workflow and Systems

These tools do not need to be built from scratch. Two repository software packages developed within the academic community are Fedora Commons[29] and DSpace[30]. Both technologies are suitable to use as a long-term digital preservation repository. In addition, there are commercial packages that can be used as a backend to a repository, including Documentum and MarkLogic. A number of collaborative projects have also leveraged the LOCKSS[31] software to develop private LOCKSS networks for preservation.

[29] http://fedora-commons.org/

[30] http://www.dspace.org/

[31] http://lockss.stanford.edu/lockss/Home

When selecting technology some important considerations are:

- Will the software meet your input throughput needs?
- Will the software meet your output throughput needs?
- How complicated is the software to manage? Do you have appropriate staff to both install the software and maintain it over time?
- Will the software capture the preservation metadata you have identified as necessary in your policies?
- Can the software support maintaining the original master versions of your content files and the web-ready versions of your content files side-by-side with the metadata for the files?
- Can the software export the original master version of your content files with the metadata for those files?
- How much does the software cost initially? Consider both internal costs such as staff time and external costs.
- How much will it cost to maintain? Consider both internal costs such as staff time and external costs such as licensing fees.

Community Monitoring

Whether an organization chooses emulation or migration (or both) as their preservation strategy, it must monitor the state of at least:

» the community of file format experts	to understand when the formats of files in the archive are reaching the end of their useful life, in order to make decisions on how to manage the situation (migration, emulation, or no action)
» the community of preservation experts	to understand what new tools are available and can be leveraged
» their content provider community	to understand whether or not the needs of the content provider community continue to be met by the archive
» their designated community	o understand whether or not the needs of the user community continues to be met by the archive

Participation in community preservation efforts brings much value to organizations. For example, many preservation standards such as PREMIS and METS are community efforts and by participating in those communities each individual organization lessens its own burdens while improving the efforts of the whole. In addition to standards, there are a number of community supported tools such as PRONOM[32], an online registry of technical information, and JHOVE[33], a file characterization tool.

Preservation Policies and Review Process

Developing preservation policies is one aspect of the "Preservation Planning" activities of the model. Each organization should develop a preservation policy review process that identifies how frequently the different policies must be reviewed and which positions within the organization have the responsibilities to review and approve each document.

Preservation policy documents should include at least the following sections:

Title:
- The title should be explicit and clear. It is advisable for the file name to match the title.

Policy Statement:
- The high level policy statement is best left general, so that it does not need to be updated frequently. It does not need to be lengthy.

Implementation Examples:
- In this section, some examples of the decisions made based upon this policy should be provided.
- By separating the Policy Statement from the Implementation Examples, it will be possible to update the examples section without reapproving the entire policy document.

Document History:
- The policy should contain a history of the versions of the document, the changes made to each version, the author of the changes, and should indicate which versions of the document are formally "approved" via the approval process at the organization.
- In addition to the document history, each document should include:
  - Who approved the policy document
  - When the policy document was last approved
  - Who reviewed the policy document

An example template is available at *Appendix: Template Preservation Policy* and a downloadable template in Microsoft Word is available at the Portico website.[34] This template is provided for guidance and as needed, organizations should modify, expand, or reduce the number of sections.

In general, preservation policies are an expansion of the topics addressed in

[32] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[33] http://www.jhove2.org/

[34] http://www.portico.org/digital-preservation/wp-content/uploads/2011/03/NEH-White-Paper-on-Preserving-Digital-Content-Preservation-Policy-Template.docx

Pre-Preservation Analysis & Planning. In addition, there may already be policies within an organization that can be leveraged while developing preservation policies and it is logical "to try to identify what exists in your organization in terms of high level policies and schedules, for example policies on financial, staffing or risk assessment."[35] A starting set of policies or topics to have within a single policy are:

- Required metadata: What metadata is required by the archive for the collection?
- Content selection: How is content selected for preservation?
- Securing preservation rights: How will rights be secured? How will rights be indicated within the preserved content?
- Modification and deletion of preserved content policy: Will modification or deletion of preserved content be allowed? If so, when and how?
- Collection specific preservation policies: Are any policies necessary for specific collections?
- Provider initiated update policy: If the original content provider has an update to make, will that be allowed? How will it be implemented? Will the update be allowed to overwrite the original content? If not, how will versioning be managed?
- Designated community and feedback policy: Who is the community and how will they be involved in the preservation service?
- Documentation and policy review cycle: What is the policy for reviewing documentation and policies?
- Migration and emulation policies: Will the archive rely on migration or emulation or both to maintain the content over time? If so, how?
- Hardware and software lifecycle and refreshment policy: How frequently will hardware and software be replaced?
- Identifier usage policy: What identifiers are used within the content? What unique identifiers will the repository assign?
- Problem resolution escalation path: When problems are discovered in the content by staff or users, how will they be escalated and resolved?
- Public disclosure of agreements policy: Will license agreements and other agreements be disclosed?
- Software development and content processing quality control policies: What quality control and assurance processes are in place?
- Replication and backup policies: What are the policies on the number of replicas, backups, and their frequency of creation, update and testing?
- Roles and responsibilities: What roles existing within the organization and what responsibilities do they have?
- Succession or end of life policies: Should the organization no longer wish to support the collection, is there a plan for where it will go?

[35] Beagrie, Neil, Najla Semple, Peter Williams, and Richard Write. "Digital Preservation Policies Study - Part 1: Final Report October 2008." Pg. 13. http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf. Last accessed on Mar 29, 2010.

Descriptive Metadata, Preservation Metadata, and Packaging

The digital building blocks of any given intellectual unit are files and usually a disparate collection of files. The following could represent a digitized book:

Figure 10: Files in a Digitized Book

It is the descriptive metadata, the preservation metadata, and the packaging metadata that allow people and machines to make sense of this mass of files.

Descriptive Metadata: There are many appropriate descriptive and full-text metadata formats for digitized content. For digitized books, common descriptive metadata formats are:

Full-Text	Header
» NLM—NCBI Book Tag Set[36] » TEI—Text Encoding Initiative[37]	» NLM—NCBI Book Tag Set » MARC[38] » ONIX for Books[39]

It is not unusual for organizations with digitized books to have both a header (or descriptive metadata only) metadata file and a full-text file.

Organizations should select one or more standard descriptive metadata formats that allow for robust expression and characterization. In addition, a format that has a large set of existing tools is desirable. Every archival information package should have a full and robust descriptive metadata file associated with it.

An abbreviated version of the descriptive metadata should be placed in the preservation metadata file or within the metadata structures of the repository. Many content management

[36] http://dtd.nlm.nih.gov/book/

[37] http://www.tei-c.org/index.xml

[38] http://www.loc.gov/marc/

[39] http://www.editeur.org/83/Overview/

and repository systems only allow for Dublin Core metadata[40] and this is an appropriate set of metadata to store within the preservation metadata file:

1. Contributor
2. Coverage
3. Creator
4. Date
5. Description
6. Format
7. Identifier
8. Language
9. Publisher
10. Relation
11. Rights
12. Source
13. Subject
14. Title
15. Type

Packaging and Preservation Metadata: One of the most important results of preserving content is making explicit the relationships among the files that are used to create an intellectual unit. The packaging and preservation metadata file is the table of contents to the intellectual unit and a key element of the archive information package.

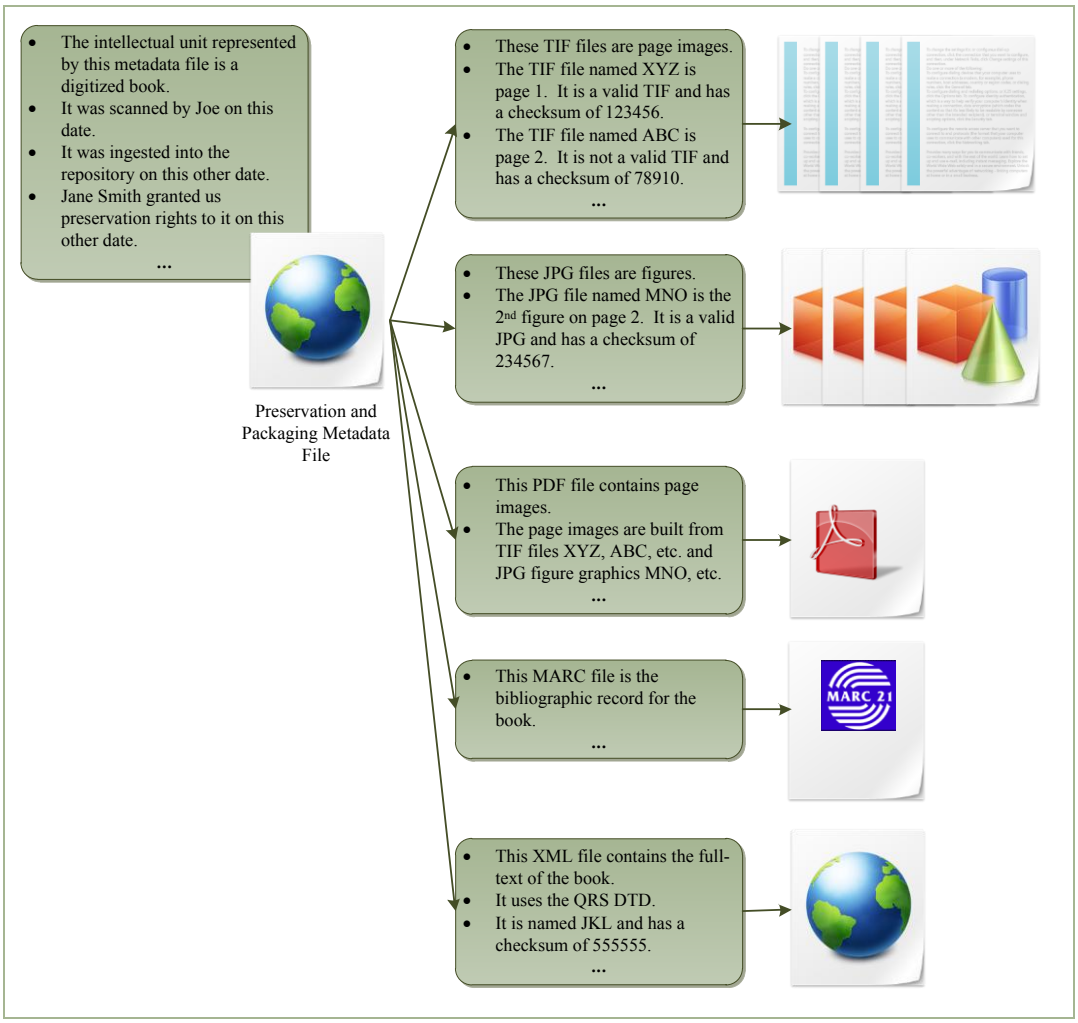


Figure 11: Depiction of the Purpose of the Packaging and Preservation Metadata File

[40] http://dublincore.org/documents/dces/

The packaging metadata is typically an XML file that describes the purpose of all the individual elements that make up the intellectual unit, their names, and how they relate to one another. If an organization chooses an off-the-shelf repository system such as DSpace or Fedora, it will most likely be committed to the packaging and preservation metadata format supported by that repository.

Within the library and institutional repository communities, METS[41] (Metadata Encoding & Transmission Standard) is the most popular structure for representing packaging. It is an XML schema that is a standard for encoding descriptive, administrative and structural metadata regarding objects within a digital library. If organizations choose to use other packaging structures (for example, DIDL, the Digital Item Declaration Language of MPEG-21,[42] or ORE, the Object Reuse and Exchange specification for the Open Archives Initiative[43]) they may wish to ensure they can transform from their internal system to METS, as this capacity may facilitate delivery to end users and replication partners.

Usually, the preservation metadata for an archival information package is included within the packaging metadata file. For example, within a METS XML file, you can include a variety of different preservation metadata schemes. Organizations should start by reviewing the PREMIS data dictionary for preservation metadata[44] and identifying how those elements can be captured within their workflow and content management systems. Organizations may choose to expand upon the PREMIS list.

If at all possible, organizations should apply a single format for packaging and preservation to all content under the organization's management. This allows the organization to manage all of its content in the same way.

Content Formats

Whether or not to define a limited set of content formats that will be allowed in the archive, or to allow all files into the archive, is a decision each organization must reach on its own. The more limited the set of formats allowed, the easier the collections will be to manage over the long-term. That ease must be weighed against the capacity of the community to invent new formats and new ways of using old formats and what willingness the organization has in limiting such innovation. This conundrum could be addressed through the use of file level preservation policies, such that the organization promises no more than byte level protection of files in more obscure formats.

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

[42] http://en.wikipedia.org/wiki/Digital_Item_Declaration_Language

[43] http://www.openarchives.org/ore/1.0/toc.html

[44] http://www.loc.gov/standards/premis/

Rights

It is important to note that simply because an organization has the right to provide access to content, does not mean they have the right to preserve the content. One step an organization should make in the content deposit process is to confirm that the entity submitting the content for preservation has the right to do so. Once preserved, the cultural heritage organization is advised to preserve the record of those rights within the archive (for example, if the organization has license agreements with content providers, those license agreements should be preserved within the archive).

Organizations should also be cognizant of several especially concerning areas of rights:

- Images, audio, and other media contained within other objects—it is possible that the content owner does not have preservation rights to content embedded within or associated with the publication.
- Privacy rights—some cultural heritage organizations may need to preserve research that contains identifying information or medical information and in such cases must be extremely cautious about privacy rights.

The Portico license agreements are publicly available and may be of some assistance as organizations consider preservation rights.

- Portico e-journal preservation license agreement[45]
- Portico e-book preservation license agreement[46]
- Portico d-collection preservation license agreement[47]

Costs

Providing long-term protection preservation and access to digital content involves costs and there is the opportunity for those costs to be substantial. Cornell University has estimated the costs of supporting arXiv.org, a widely used preprint service targeted at the physics community, at $400,000 a year[48].

As one element of the preservation planning, cultural heritage organizations will determine how long content must survive. With that decision in hand, organizations can then estimate the costs to maintain the content for that length of time and then must consider ways to meet those costs.

---

[45] http://www.portico.org/digital-preservation/wp-content/uploads/2009/12/E-Journal-License-Agreement-v.-3.3.pdf

[46] http://www.portico.org/digital-preservation/wp-content/uploads/2009/12/E-Book-License-Agreement-v.3.3.pdf

[47] http://www.portico.org/digital-preservation/wp-content/uploads/2009/12/D-Collections-License-Agreement-v.3.3..pdf

[48] See the arXiv Support FAQ at http://arxiv.org/help/support/faq, last accessed on Mar 31, 2011

Organizations should at least consider the following types of start-up costs:

- **Staff Costs**: Each organization will need to determine the skills of the staff necessary for their preservation project. Each of those staff will have a yearly salary from which a weekly salary can be computed and multiplied with the number of weeks of work that person will need to accomplish. When doing this analysis, consider whether the appropriate yearly rate is:
  - Salary alone
  - Salary plus benefits
  - Salary plus benefits plus overhead
- **Hardware and Software Costs**: In general, there will be a set of one-time costs for purchasing hardware and software.

Organizations must also consider the ongoing, annual costs of maintaining the preservation service:

- **Staff Costs**: As described elsewhere in this document, a preserved set of content will require ongoing collection and content management.
- **Hardware Replacement Costs**: We recommend that enterprise class servers be replaced every five years. For commodity Intel based servers, we recommend replacing them every three years in order to take advantage of the improvements in performance and stability which could potentially reduce costs by reducing the footprint in the datacenter. Most disks will come with a recommended replacement time frame, typically at three years. These replacement costs may be amortized over the years between replacements.
- **Annual Hardware and Software Costs**: Just as there are staff costs that recur every year, so too are there hardware and software costs that recur. They may recur yearly or perhaps monthly.

A worksheet to help calculate these costs is available in *Appendix: Worksheet to Estimate Costs*. In addition, the worksheet is available in Excel format on the Portico website.[49]

[49] http://www.portico.org/digital-preservation/wp-content/uploads/2011/03/NEH-White-Paper-on-Preserving-Digital-Content-Preservation-Costs-Worksheet.xlsx

Models for Recovering Costs

The Cornell University Library FAQ on arXiv support[50] developed a list of models for recovering preservation costs based heavily upon the ITHAKA S+R report, "Sustainability and Revenue Models for Online Academic Resources"[51]:

- Allowing sponsorship of the collection
- Permitting advertising on the collection
- Encouraging donations to support the collection
- Building an endowment
- Creating premium services for purchase, the revenues from which can subsidize the preservation service.
- Enlisting support from funding bodies, scholarly and professional societies, and publishers

Cultural heritage organizations could consider additional models, including:

- Charging for access to the collection and using the revenue to subsidize the preservation costs.
- Charging for participation in the preservation service.
- Relying upon support from a parent organization or government.

[50] http://arxiv.org/help/support/faq

[51] http://www.ithaka.org/ithaka-s-r/strategyold/sca_ithaka_sustainability_report-final.pdf

Section IV. APPENDICES

13. APPENDIX: GLOSSARY

Note that this glossary contains terms both found within this white paper and terms likely to be encountered by cultural heritage organizations as they perform their own research into the topic of digital preservation.

AHDS: The UK Arts and Humanities Data Service – funding was withdrawn in April 2008 and some services were taken over by Centre for e-Research (CeRch) at King's College London.

ALTO: An acronym for Analyzed Layout and Text Object – it is an XML schema that supports encoding OCR-recognized text and the position of that text on the source image at the word level. It is often encoded within METS and in such instances it is referred to as METS/ALTO.

ARL: The Association of Research Libraries.

Artesia: A commercial digital asset management system from Open Text.

CDL: California Digital Library – founded by the University of California to take advantage of emerging technologies.

CEN.BT TF 179: A shorthand notation for the Cinematographic Works Standard metadata framework being created under the auspices of the Task Force 179 of the European Committee for Standardization (CEN). It has been since superseded by CEN.BT Technical Committee 372.

CMS: A content management system – it is software designed to allow organizations to manage their digital objects. It sometimes has a hardware component, as well as a software component.

CONTENTdm: A digital repository system from OCLC – it is available both as a local installation and as an OCLC hosted service and is most frequently used as a hosted service.

Copac: A freely available library catalogue with approximately 32 million records and representing the merged holdings of the members of the Research Libraries UK (RLUK) - this includes the catalogues of the British Library, the National Library of Scotland, and the National Library of Wales / Llyfrgell Genedlaethol Cymru and increasing numbers of specialist libraries with collections of national research interest.

DAMS: An acronym for "Digital Asset Management System" – it is being built at Oxford University to provide long-term content management to digital content.

DC: See Dublin Core.

Digital Preservation: the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long-term. The key goals of digital preservation include: usability – the intellectual content of the item must remain usable via the delivery mechanism of current technology; authenticity – the provenance of the content must be proven and the content an authentic replica of the original;

discoverability – the content must have logical bibliographic metadata so that the content can be found by end users through time; and accessibility – the content must be available for use to the appropriate community.

Digital Repository System: Software to enable the collection of content on the web – they are similar to content management systems, but do not enable the creation of robust content management workflows.

DLS: Also known as Digital Library System – it is the software that has been built by the British Library to provide itself with long-term digital preservation.

DMD: An abbreviation for descriptive metadata – it is bibliographic metadata that describes an object.

Drupal: An open source content management platform developed for website maintenance.

DSpace: An open source digital repository package.

DTD: A document type definition – it is a specific definition that follows the rules of the Standard Generalized Markup Language (SGML and provides a specification that accompanies a document and identifies markup elements and the rules for their use.

Dublin Core: A shorthand notation for the "Dublin Core Metadata Element Set", which is a vocabulary of fifteen properties for use in resource description. It is abbreviated, DC.

EAD: The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids using XML.

EDINA: EDINA is the JISC national academic data centre based at the University of Edinburgh – it has a mission to enhance the productivity of research, learning and teaching across all universities, research institutes and colleges in the UK.

Extensis Portfolio: A commercial digital image management system to allow for cataloging of files, visual organization of files, and drag and drop integration with the operating system.

FE: An abbreviation for Further Education.

Fedora: It is an open source content management platform that enables the storage, access and management of digital content

GB: An abbreviation for gigabyte – it is 1,000,000,000 bytes or 109 bytes. A project with content in the gigabytes is relatively small.

GIS: An abbreviation for geographic information system – it is a system that captures, stores, analyzes, manages, and presents data that is linked to location. GIS is often used to refer to the data that drives a geographic information system.

HD: An abbreviation for high definition.

HE: An abbreviation for Higher Education.

HFS: An abbreviation for Hierarchical File System – which is a robust file server and backup system maintained by Oxford University Computing Services.

ISAD(G): A standard that provides general guidance for the preparation of archival descriptions. It is used in conjunction with existing national standards or as the basis for the development of national standards.

JISC1: An acronym used for projects that received funding through Phase 1 of the JISC Digitisation Programme.

JISC2: An acronym used for projects that received funding through Phase 2 of the JISC Digitisation Programme.

JORUM: A free online service providing access to teaching and learning resources, for teaching and support staff in UK Further and Higher Education Institutions

JPG: Also abbreviated as JPEG (Joint Photographic Experts Group) – it is the file ending of images using the JPEG method of compression and is often used as a shorthand notation for files of this type.

LTO: An abbreviation for Linear Tape-Open – it is an open standard magnetic tape data storage technology.

MARC: A library standard format for the representation and communication of bibliographic and related metadata in machine-readable form.

MARCXML: A framework for working with MARC data in a XML environment.

MD: An abbreviation for metadata – data that describes other data or content.

METS: Metadata Encoding and Transmission Standard -- The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium

MIMAS: A JISC and ESRC-supported national data centre providing the UK Higher Education, Further Education and research community with access to key data and information resources to support teaching, learning and research across a wide range of disciplines.

MINISIS: A commercial archive collection management software package.

MIX: An XML schema for a set of technical data elements required to manage digital image collections. The schema provides a format for interchange and/or storage of the data specified in the Data Dictionary - Technical Metadata for Digital Still Images (ANSI/NISO Z39.87-2006). This schema is currently referred to as "NISO Metadata for Images in XML (NISO MIX)."

MODES: A shorthand notation for MODES Catalog System, which is an old cataloging system designed for special collections and in use by several of the JISC digitisation projects.

MODS: The Metadata Object Description Schema – it is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. It can carry selected data from existing MARC 21 records as well as enabling the creation of original resource description records.

MP3: An abbreviation for MPEG-1 Audio Layer 3 – it is a digital audio encoding format using a form of lossy data compression. It is a common audio format for consumer audio storage, as well as a de facto standard encoding for the transfer and playback of music on digital audio players. (It should not be confused with MPEG-3 which is a group of audio and video coding standards agreed upon by the Moving Picture Experts Group (MPEG) designed to handle high-definition television signals).

OAI-ORE: The Open Archives Initiative Reuse and Exchange protocol. It defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation.

OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting. It is a mechanism for repository interoperability. It assumes data providers that are repositories which expose structured metadata via OAI-PMH and service providers that make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.

OCR: It is an abbreviation for optical character recognition, which is the recognition of printed or written text characters by a computer. The term OCR is often used to label the text files created through optical character recognition.

OUCS: An abbreviation for Oxford University Computing Services.

OULS: An abbreviation for Oxford University Library Services.

PCM: Pulse Code Modulation – it is the usual bitstream encoding format used for WAV files.

PNG: An abbreviation for Portable Network Graphics – it is a bitmapped graphics file format endorsed by the World Wide Web Consortium and is expected to eventually replace the GIF format. PNG provides advanced graphics features such as 48-bit color, including an alpha channel, built-in gamma and color correction, tight compression and the ability to display at one resolution and print at another.

Portfolio: See Extensis Portfolio.

PREMIS: An acronym used to represent the elements of the PREMIS Data Dictionary for Preservation Metadata.

PRINCE2: PRINCE2 is a generic project management method that covers how to organise, manage and control projects.

QA: An abbreviation for quality assurance – it is often used as a shorthand notation for the staff who perform quality assurance on a project.

RLUK: An abbreviation for Research Libraries UK.

SPECTRUM: A UK and international standard for collections management – it is used by museums and other cultural heritage organizations. It includes a standard format for

exchanging object records between different Collections Management Systems, support for rights management, and support for the exchange of User Generated Interpretation through the Revisiting Collections methodology.

TB: An abbreviation for terabyte – it is 1,000 gigabytes.

TEI: A shorthand notation for a set of guidelines created by the Text Encoding Initiative, which is a consortium that collectively develops and maintains a standard describing encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.

textMD: A XML Schema that details technical metadata for text-based digital objects. It most commonly serves as an extension schema used within the METS administrative metadata section

TIFF: Tagged Image File Format (abbreviated TIF or TIFF) is a file format for storing images, including photographs and line art.

TMD: An abbreviation for technical metadata – it is metadata that describes the technical format of an object.

UKDA: An abbreviation for the UK Data Archive – it is a centre of expertise in data acquisition, preservation, dissemination and promotion; and is curator of the largest collection of digital data in the social sciences and humanities in the UK.

VITAL: A commercial institutional repository product from VTLS and built on Fedora.

WAI: A shorthand notation for the best current practice for embedding accessibility roles and states in HTML documents as defined by the Web Accessibility Initiative (WAI) Protocols and Formats working group.

WAV: An acronym for the Waveform audio format (also abbreviated as WAVE) – it is a Microsoft and IBM audio file format standard for storing an audio bitstream.

14. Appendix: Participants in the Portico Locally Created Content Study

1. Baylor University
2. Binghamton University
3. Brigham Young University
4. California State Polytechnic, Pomona
5. Case Western University
6. City University of New York
7. Colorado State University
8. McMaster University
9. Middlebury College
10. Northwestern University
11. Queensland University
12. Trinity College Dublin
13. University of British Columbia
14. Vassar College

15. Appendix: Participants in the JISC Preservation Study

1. Birmingham Museums & Art Gallery, Pre-Raphaelite Resource Site
2. Bournemouth University, Digitisation of the Independent Radio News (IRN) Archive
3. British Film Institute, InView: Moving Images in the Public Sphere
4. British Library, Archival Sound Recordings 2
5. British Library, British Newspapers 1620-1900
6. British Library, UK Theses Digitisation Project
7. Cambridge University, Freeze Frame
8. The National Archives, Cabinet Papers, 1915-1978
9. National Library of Wales, Welsh Journals Online
10. Oxford University, First World War Poetry Archive
11. Oxford University, The John Johnson Collection: An Archive of Printed Ephemera
12. Queen's University at Belfast, A Digital Library of Core E-Resources on Ireland
13. University of East London, The East London Theatre Archive (ELTA)
14. University of Kent, British Cartoon Archive Digitisation Project
15. University of Portsmouth, Historic Boundaries of Britain (HBB)
16. University of Southampton, 19th Century Pamphlets Online

16. Appendix: Straw-man Description of Possible Portico Preservation Service for Locally Created Content (LCC)

LCC Straw-man

PORTICO

Straw-man Description of Possible Preservation Service for Locally Created Content (LCC)

Portico is investigating offering a service to libraries and other cultural heritage organizations to preserve their locally created digital content. Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. Through in-depth discussions with 16 libraries involved in the study, we have developed the following straw-man preservation service description.

Preservation should be tailored to the needs of the "designated community" for whom the content is preserved, and, thus, each institution participating in any Portico LCC preservation service would work with Portico to define its own preservation service through configuration of the ingest and migration phase of the preservation process. Ingest and migration is the second phase of Portico's four phases of operational activities.

Phase 1: Transfer to Portico: Move a copy of the content from the institution to Portico.

Phase 2: Ingest into Portico archive and Migration: This phase of preservation activities may be configured by participating institutions. The choices include:

- *Full Preservation Activities*:
  - Analyze the structure of the content to determine whether all expected files were received
  - Validate files against their format specifications and revalidate files in the future as new tools are developed
  - Repackage content into a Portico archival information package (AIP)
  - Migrate files to new formats on ingest or in the future as necessitated by the changing technological environment
- *Full Preservation Activities plus Text Migration*:
  - Apply the above full suite of preservation activities
  - Transform descriptive metadata or structured full-text to a standard format
- *Full Preservation Activities without Local Objects*:
  - Apply the full suite of preservation activities to content that is not held within the Portico archive – Portico will maintain the preservation metadata and perform migrations, but will not retain any original content files, which will instead be retrieved in the future from your institutional data store
- *Zip and Hold*:
  - Package content into ZIP files and hold it in the Portico archive, with preservation activities limited to standard archive maintenance (see below)

Phase 3: Portico Archive Maintenance: Standard Portico archive maintenance includes on- and off-line replication, on- and off-line media refreshment, fixity and completeness checks, receipt and processing reports, audit accreditation reports, and monthly status reports on holdings, repairs, fixity, completeness and migrations.

Phase 4: Dissemination from Portico: Standard Portico archive dissemination includes audit access for select institution staff and export of the content back to the institution in original or archival packaging.

Page 1 of 4

LCC Straw-man

PORTICO

Portico Operational Activities and the LCC Preservation Service Straw-man Description

❶ Transfer to Portico

- **INITIAL**: crawl of specified collections per agreement
- **UPDATES**: monthly crawls for updates, deletions, and additions

FROM ONE OF THE FOLLOWING SOURCES

- Fedora or DSpace with Portico plugin
- Other repository with implementation of Portico OAI-PMH specification
- Other repository with OAI-ORE

A limited preservation service is available for institutions *without* OAI-ORE or an implementation of OAI-PMH to the Portico specifications. This limited service includes transfer to Portico of the data by FTP or delivery on media or by a Portico initiated web crawl of the institutional website. It does not include any provision for institution originated updates, deletions, or additions to the content. The only ingest and migration option for content delivered in this manner is "zip and hold."

❷ Ingest into Portico Archive and Migration

- Full preservation activities: structural analysis, file format validation, repackaging of content into archival packaging, and migration of object files

... OR ...

- Full preservation activities plus text migration: above suite of activities, plus migration of descriptive metadata or structured full-text

... OR ...

- Full preservation without local objects: above suite of activities without retention of original content files - a URI to the location on your institutional data store remains with the Portico preservation metadata

... OR ...

- Zip and Hold: content is packaged into ZIP files and retained in the Portico archive

❸ Portico Archive Maintenance

- On & off-line replication and media refreshment
- Fixity and completeness checks
- Audit accreditation reports
- Receipt & processing reports and monthly status reports on holdings, repairs, fixity, completeness, and migrations

❹ Dissemination from Portico

BASIC SERVICE

- Audit access for select institution staff
- Export to institution of original content in original packaging
- Export to institution of original and migrated content in Portico AIP packaging

ADDITIONAL OPTIONS

- Trigger Event - temporary
- Trigger Event - permanent
- Ongoing access

Page 2 of 4

LCC Straw-man

PORTICO

Example Configurations of the LCC Preservation Service

Full Preservation Activities	Full Preservation Activities Plus MD Migration	Zip & Hold
Example: a DSpace installation with the Portico plug-in. ○ Portico does an initial content transfer, validates the files to their format specifications, analyzes the structure to ensure that all files have been received, and repackages the content. ○ Receipt and processing reports are sent to the institution. ○ Portico queries DSpace for additions, deletions, and updates on a monthly basis, and after ingest of the updates a receipt and processing report is sent to the institution. ○ Portico sends the institution monthly status reports on their holdings. ○ Portico validates files to their format specification as tools are developed and will migrate object files as necessary. ○ Audit access to the content and delivery to the institution upon request.	Example: a home grown content management system (CMS) with an OAI-PMH interface built to the Portico specifications (to allow Portico to retrieve content files). The CMS contains e-books in a consistent format and MARC metadata records. ○ Portico transfers content, validates the files, analyzes the structure to ensure that all files have been received, migrates the e-books plus their MARC data to the NLM e-book standard and a qualified Dublin Core. (This step is configurable based upon content and institution needs). ○ Receipt and processing reports are sent to the institution. ○ Portico will query the CMS for additions, deletions, and updates on a monthly basis, and after ingest of the updates a receipt and processing report is sent to the institution. ○ Portico sends the institution monthly status reports on their holdings. ○ Portico validates files to their format specification as tools are developed and will migrate object files in the future, including the metadata and e-book files. ○ Audit access to the content and delivery to the institution upon request.	Example: a digital collection run on CONTENTdm without a content transfer implementation. ○ Portico will crawl the website. The web pages and images for each object will be zipped together into a single unit and placed in the archive. ○ Receipt and processing reports are sent to the institution. ○ Portico sends the institution monthly status reports on their holdings. ○ Audit access to the content and delivery to the institution upon request.

Page 3 of 4

What is Digital Preservation?

Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. The key goals of digital preservation include:

- usability – the intellectual content of the item must remain usable via the delivery mechanism of current technology
- authenticity – the provenance of the content must be proven and the content an authentic replica of the original
- discoverability – the content must have logical bibliographic metadata so that the content can be found by end users through time
- accessibility – the content must be available for use to the appropriate community

17. Appendix: Template Preservation Policy

<Preservation Policy Title>

1. Policy Statement

1.1. <Paragraph 1 in high level policy statement>

1.2. <Paragraph 2 in high level policy statement. This section should be short, but accurate and to the point. It should provide guidance to operations staff as they do their jobs.>

2. Implementation Examples

2.1. <Example 1>

2.2. <Example 2 – examples should describe practical decisions made based upon this policy.>

3. Document History

3.1. Approved by: <Name of approver>

3.2. Last Review Date: <Date policy was last approved>

3.3. Reviewed by: <Names of people who reviewed the document at ita last review date>

3.4. Change history:

Version	Date	Change	Author
<version #>	<date finalized>	<Highlight the changes made to the document>	<author of the changes>
<version 2>*	<date finalized>	<Highlight the changes made to the document>	<author of the changes>

* An approved version of this document.

A Microsoft Word version of this template is available on the Portico website.[52]

[52] http://www.portico.org/digital-preservation/wp-content/uploads/2011/03/NEH-White-Paper-on-Preserving-Digital-Content-Preservation-Policy-Template.docx

18. Appendix: Illustrations of Answers to the Practical Questions

Note that these illustrations were originally written as part of a report to JISC.

Project A

A University department has a special collection of primary source documents that has grown over time. The department hires a project manager for the duration of the creation of the digital collection to coordinate the digitization of the content and creation of descriptive metadata. With the help of the University IT department, they place the content into an institutional repository and make it available for use on-line.

Project A - Digital Preservation Policies and Plan

Who: Identify the key players involved with long-term preservation of the targeted content.

Who is writing the policy and plan?

> The digitization project manager in the Department of Lake Studies at the University of Lorem Ipsum.

Who will use the content in the short and long-term?

> The content should be made available for use by anyone in the world.

Who has responsibility for maintaining the intellectual content of this collection (e.g. making corrections to metadata or content files)?

> The University of Lorem Ipsum IT department has responsibility for the ongoing maintenance of the collection in the institutional repository. If corrections are suggested through user feedback, the IT department should contact the Department of Lake Studies administrator who will then discuss the correction with the Department Director and approve or disapprove it. The IT department will make the changes in the institutional repository.

Who has responsibility for maintaining the bytes of the files in this collection (e.g. identifying and fixing corrupted files)?

> The University of Lorem Ipsum IT department will ensure that the content files do not become corrupted.

Who approved this policy and plan?

Ms. Jones, Director of the IT Department, University of Lorem Ipsum

Mr. Challa, Director of the Department of Lake Studies, University of Lorem Ipsum

What: Describe or characterize the collection and content.

What is the content and from where did the content originate?

The content is digitized postcards, letters and other ephemera. A large portion of it was donated to the Department in 1965 by Mr. Smith. The faculty of the Department of Lake Studies has added to the collection since that time.

What file formats, including metadata formats, are present?

The content has been digitized as TIFF images (300 dpi, 48 bit color). The descriptive metadata is first captured in the Department of Lake Studies catalog (which is used to describe the analog content in the collection, as well). The images are referenced by filename in the catalog record. The catalog records and TIFF images are exported to the IT department and are placed in the institutional repository.

How many items are in the collection? How large is the collection on disk?

There are 4000 images in the collection covering 1000 postcards, 2500 letters, and other ephemera. It is approximately .5 TB.

Where: Document the locations of all the copies of the content and metadata.

Where is the master copy of the descriptive metadata kept?

The master copy of the descriptive metadata is kept in the Department of Lake Studies catalog.

Where is the master copy of the content files kept?

The master copy of the content files is kept in the institutional repository maintained by the IT department.

Where are all the copies of the content, including backups, and how are the copies of the content related?

> The institutional repository also has a copy of the metadata, however it is a derivative and not as robust as what is held in the Department of Lake Studies catalog. The catalog has monthly full backups and weekly incremental backups that are housed in the IT department's machine room. The institutional repository also has monthly full backups and weekly incremental backups. In addition, it has monthly backups to tape which are sent off-site.

When: Document the targeted preservation timeframe and impact of loss.

How long should the content be available for use?

> The content should remain available for use for at least 50 years.

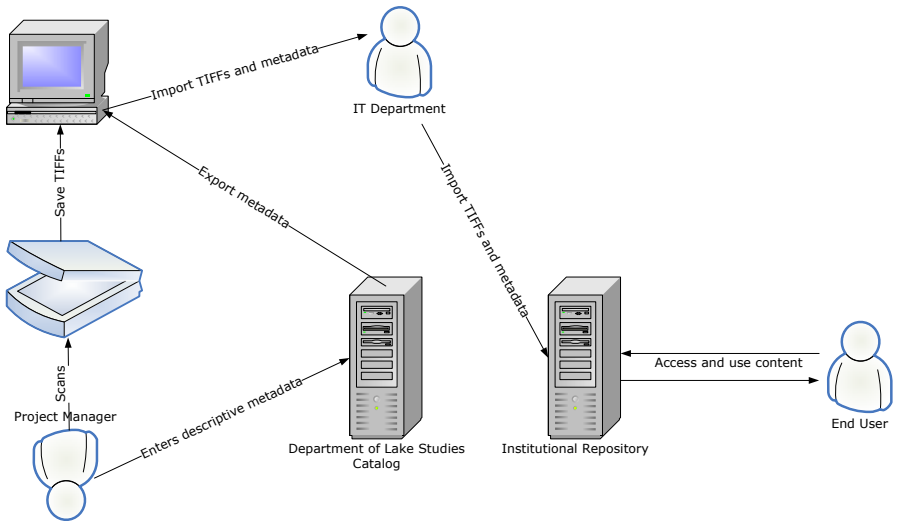
If the content is irretrievably lost, what are the repercussions?

> There are none.

How: Document how the key content management and preservation tasks will occur.

How will the collection be created (perhaps draw a diagram of the workflow)?

The Project Manager or an intern in the Department of Lake Studies will scan the artifacts and create TIFF images. These are temporarily saved to a local computer. The same person will enter descriptive metadata into the Department of Lake Studies catalog. At the end of every month, the Project Manager exports the metadata from the catalog to the image file server in Excel format. The IT department captures the Excel file and the appropriate TIFFs and places them in the institutional repository. The repository turns the TIFFs into JPGs on the fly when requested by an end user. The images on the file server at the Department of Lake Studies will be deleted after they are successfully loaded into the institutional repository.



How will the collection be maintained (perhaps draw a diagram of the workflow)?

The Department of Lake Studies will not be performing regular maintenance of this digital collection. If a metadata update needs to be made, the IT department may contact the Department of Lake Studies and the changes will be made manually at both the institutional repository and the Department of Lake Studies catalog. If the Department of Lake Studies initiates a correction, they will contact the IT department to synchronize the updating of metadata.

Do you expect the content files to be migrated in the future?

If it is necessary to migrate the files within the next 50 years so that the collection remains usable, yes.

May the content files be deleted? Added to? Updated?

> The content files may be deleted and updated – though it should be rare and only in the case of an error. The collection will be closed when the project is completed and no additional content files will be added.

May the descriptive metadata be deleted? Added to? Updated?

> The descriptive metadata may be updated. It should not be deleted, though a note may be made that the digitized file(s) to which it refers has been deleted. As the collection will be closed when the project is completed, we do not anticipate entry of new metadata records.

How will you track who did what and when to the content, if this is important to your organization?

> It is not important and will not be tracked.

How do you associate the master copy of the descriptive metadata with the master copy of the content files and how will you move these two items around together?

> The metadata record in the institutional repository is not complete. The reasons for any need to ship the content must be analyzed. For certain purposes, an export from the institutional repository may be sufficient. If a master copy of the metadata must be exported with the master copy of the images, then the IT Department will need to coordinate with the Department of Lake Studies to merge the metadata in the catalog with the images in the institutional repository. There is no automatic way to do this.

Project B

A large library has digitized old analog video recordings. The analog version of the video recordings is secure in the institutional video vault. The library does not have a the skills or desire to provide access to this content and has therefore shipped a copy of the content to a third party access service that specializes in delivery of digital video recordings. That third party service has agreed to provide access to the content for at least 10 years. The third party access service creates smaller delivery files from the master copy of the content provided by the library and then deletes its copy of the original. The library is maintaining a preservation copy of the original digitized recordings in its robust, institutional archive.

Project B - Digital Preservation Policies and Plan

Who: Identify the key players involved with long-term preservation of the targeted content.

Who is writing the policy and plan?

Mr. Joe Kline, Director of Library Video Services, University of Smithtown

Who will use the content in the short and long-term?

The UK HE and FE community

Who has responsibility for maintaining the intellectual content of this collection (e.g. making corrections to metadata or content files)?

The staff of the library video services department is responsible for updating the metadata and content files within the institutional preservation service.

Who has responsibility for maintaining the bytes of the files in this collection (e.g. identifying and fixing corrupted files)?

The University of Smithtown's IT department is responsible for maintaining the institutional archive and will provide required ongoing preservation maintenance to this content.

Who approved this policy and plan?

Ms. Adelaide Bovie, Director of the Library, University of Smithtown

Mr. Muhammad Bishara, Director of Information Technology, University of Smithtown

What: Describe or characterize the collection and content.

What is the content and from where did the content originate?

The University of Smithtown has long been a center for film and video studies and over the decades, the university library has developed an extensive collection of analog videos that are now out-of-copyright.

What file formats, including metadata formats, are present?

> The videos masters are in WAV format. The metadata is in a proprietary framework and uses a qualified Dublin Core for the descriptive metadata. Submaster files, also in WAV format, are created from the original masters. These are each a clip from the original.

How many items are in the collection? How large is the collection on disk?

> There are 500 master files, with just under 5000 submasters. The collection is approximately 1.5 Tb.

Where: Document the locations of all the copies of the content and metadata.

Where is the master copy of the descriptive metadata kept?

> The master copy of the descriptive metadata is in the institutional archive.

Where is the master copy of the content files kept?

> The master and submaster copies of the content files are in the institutional archive.

Where are all the copies of the content, including backups, and how are the copies of the content related?

> The access provider has a copy of access derivatives of the content and the metadata. This is not tied back to the master copies at all – though if needed, it could be traced through original file name. The access provider is responsible for its own backups.
>
> At the University of Smithtown, the master content files, submasters and metadata are all held within the institutional archive. This service is on RAID 5 servers with a 9.99% uptime guarantee. Disk snapshots are made to an off-site, University owned machine room nightly and weekly full backups are written to tape. The tapes are kept on-site for one month and then moved off-site for storage for 3 months.
>
> Within the institutional archive, this content is all filed as the "Library Video Collection".

When: Document the targeted preservation timeframe and impact of loss.

How long should the content be available for use?

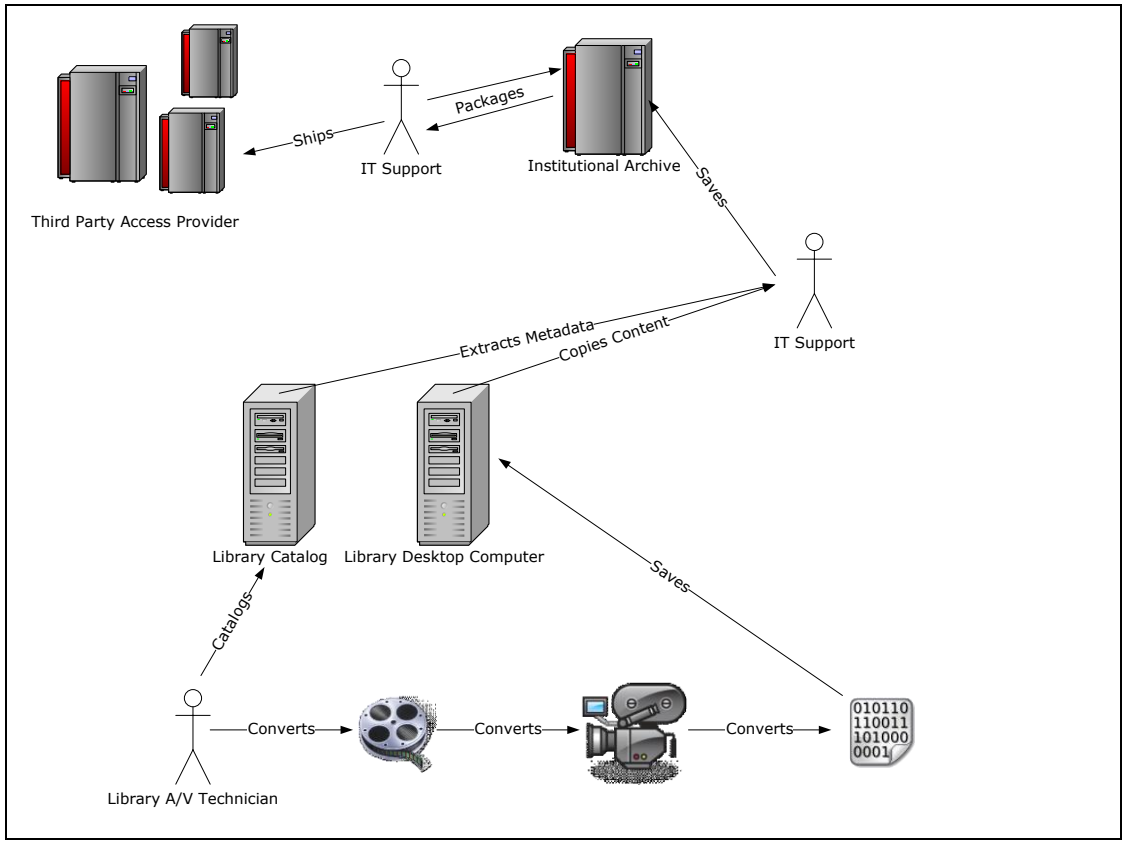
> The content should be available for use for 10 years.

If the content is irretrievably lost, what are the repercussions?

The content would have to be redigitized from the analog. This will be possible, as the analogs are secure, but it would be expensive.

How: Document how the key content management and preservation tasks will occur.

How will the collection be created (perhaps draw a diagram of the workflow)?



How will the collection be maintained (perhaps draw a diagram of the workflow)?

The IT Department does not yet have robust toolsets for editing the metadata or updating the content files within the institutional archive, however they will be developed over time. In the mean time, should metadata need to be updated it will be updated within the library catalog and then the library staff and IT staff will coordinate on updating the preserved content. The same is true if content files must be updated.

Do you expect the content files to be migrated in the future?

Yes. The IT staff has committed to this and it is expected as part of the institutional archive.

May the content files be deleted? Added to? Updated?

The content files will not be deleted from the archive, they may be updated.

May the descriptive metadata be deleted? Added to? Updated?

The descriptive metadata may be update over time.

How will you track who did what and when to the content, if this is important to your organization?

The metadata structure in the institutional archive has the PREMIS concepts of events included and an event record will be made for every update.

How do you associate the master copy of the descriptive metadata with the master copy of the content files and how will you move these two items around together?

The institutional archive keeps the two together.

Project C

A national library has a significant collection of books published in the mid-19th century on acidic paper. It is digitizing this collection in advance of the books disintegrating. The library has a content management system that will allow it to provide access to the content and is outsourcing the preservation of these digitized materials.

Project C - Digital Preservation Policies and Plan

Who: Identify the key players involved with long-term preservation of the targeted content.

Who is writing the policy and plan?

Mr. Jason Jackson, Manager of Digital Collections, the National Library

Who will use the content in the short and long-term?

The general public.

Who has responsibility for maintaining the intellectual content of this collection (e.g. making corrections to metadata or content files)?

> The Digital Collections department of the National Library.

Who has responsibility for maintaining the bytes of the files in this collection (e.g. identifying and fixing corrupted files)?

> The Third Party Preservation Service.

Who approved this policy and plan?

> Dr. Meredith Jones, Director of the National Library

What: Describe or characterize the collection and content.

What is the content and from where did the content originate?

> The content was digitized from the brittle and crumbling collection of 19th and 20th century books owned by the National Library. The library has developed a project plan which lays out the order in which different subjects and years will be digitized. Please contact the Manager of Digital Collections for further details.

What file formats, including metadata formats, are present?

> The final product is one PDF file per book with its corresponding MARC record from the library catalog.

How many items are in the collection? How large is the collection on disk?

> The collection is currently 1000 books and is approximately 500 Gb. This project is ongoing and we estimate that the library has over 35 miles of shelves of books to digitize. The project is budgeted for the next 5 years and we anticipate digitizing 500 books a year.

Where: Document the locations of all the copies of the content and metadata.

Where is the master copy of the descriptive metadata kept?

> The master copy of the descriptive metadata is in the libraries access system. While the data originated in the library card catalog, that data is *not* considered the master descriptive metadata.

Where is the master copy of the content files kept?

> The master copy of the content files is kept on the libraries "S" drive. This is also known as shareddrive-s.nationallibrary.net. The access system or content management system is run on Fedora. The large TIFFs and OCR files that are used to create the derivative PDFs are not within Fedora, but each book record in Fedora does point to the TIFF and OCR files in their home location on the "S" drive.

Where are all the copies of the content, including backups, and how are the copies of the content related?

> A snap shot of the S drive is taken nightly and placed on a machine within the same machine room.
>
> The master content files have monthly full backups and daily incremental backups to tape. The tape jukebox is held on a separate machine room off-site.
>
> The access site (which includes the master copy of the descriptive metadata) has monthly full backups and daily incremental backups to tape and to cloud storage. In addition the access site is fully synchronized with a machine in a separate machine room off-site – live fail-over can occur and has been tested.
>
> The long-term preservation of this content is being managed by the Third Party Preservation Service, which holds a complete copy of the PDFs, TIFFs, OCR, and metadata records within its fully replicated archive.

When: Document the targeted preservation timeframe and impact of loss.

How long should the content be available for use?

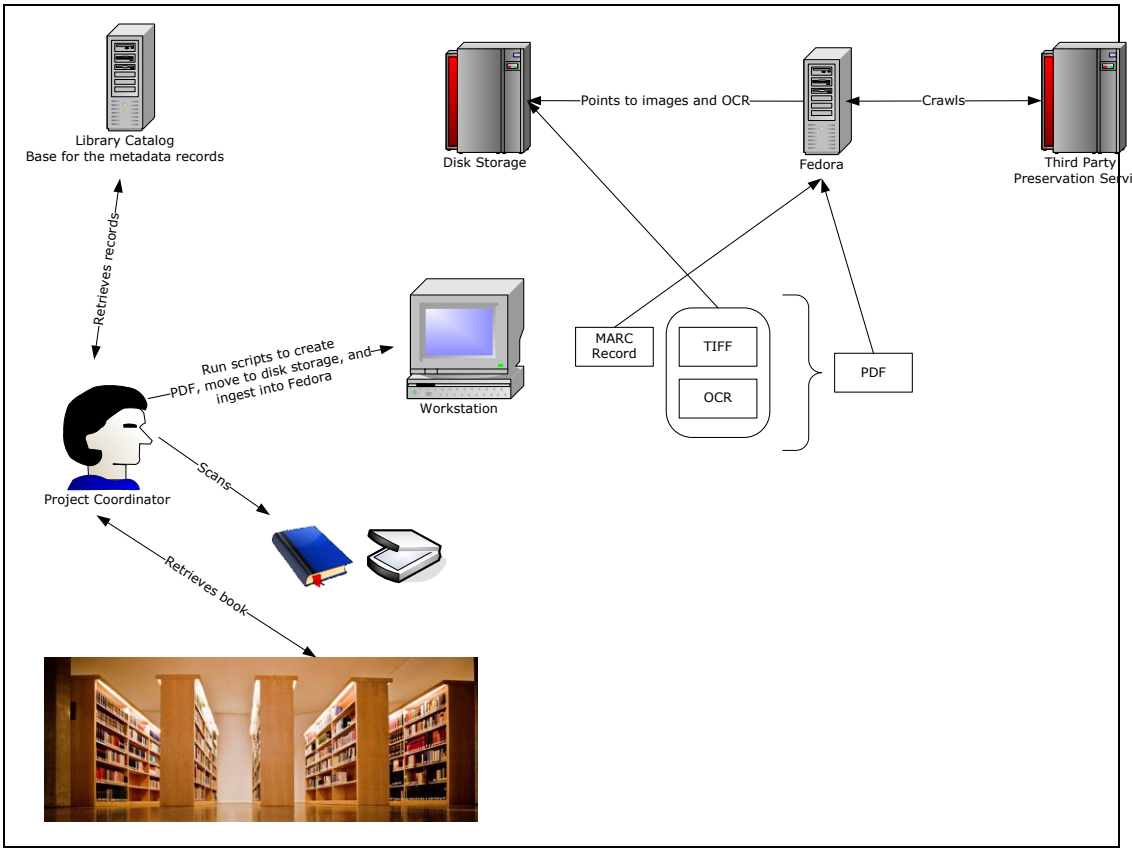
> The content should be available for use forever.

If the content is irretrievably lost, what are the repercussions?

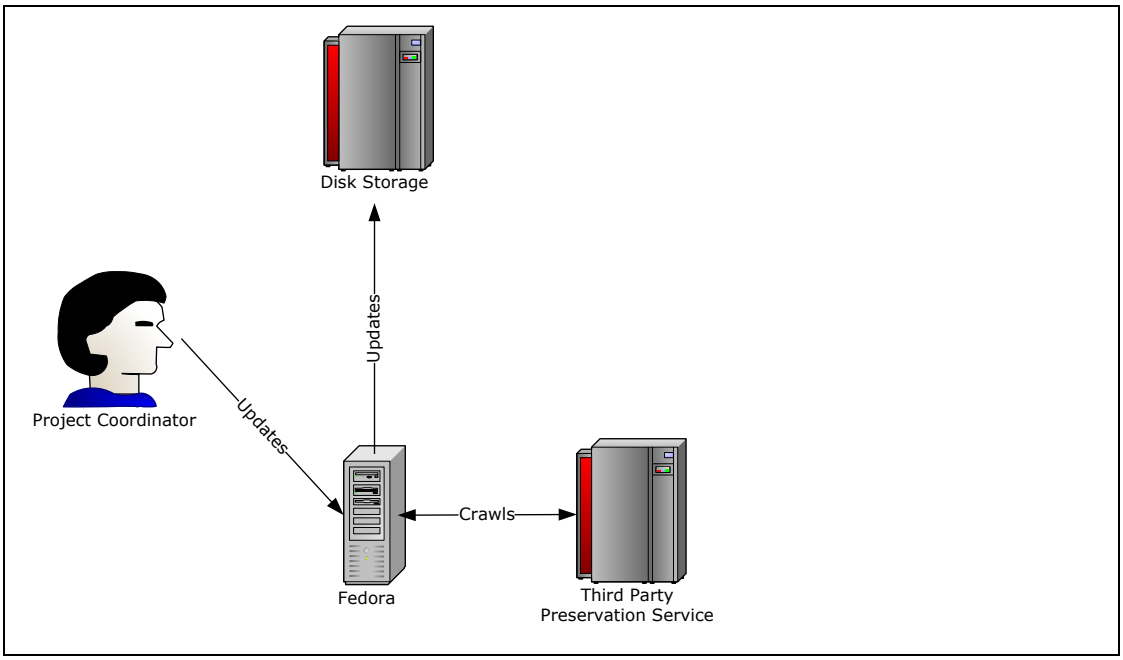
> The repercussions are large. The paper cannot be redigitized, it is too fragile. Our only copy of this content is the digital version.

How: Document how the key content management and preservation tasks will occur.

How will the collection be created (perhaps draw a diagram of the workflow)?



How will the collection be maintained (perhaps draw a diagram of the workflow)?



Do you expect the content files to be migrated in the future?

Yes, the Third Party Preservation Service will migrate the content as needed over time.

May the content files be deleted? Added to? Updated?

Content files may occasionally be deleted if its necessary for clean-up or copyright issues. The collection will grow with time.

May the descriptive metadata be deleted? Added to? Updated?

Yes, descriptive metadata may be deleted, increased, and updated.

How will you track who did what and when to the content, if this is important to your organization?

The Fedora system will track what was changed when and by whom.

How do you associate the master copy of the descriptive metadata with the master copy of the content files and how will you move these two items around together?

It is tied together both in Fedora and in the Third Party Preservation Service.

19. Appendix: Software Systems in Use Across Both Studies

Type of Software System	Specific System	Instances
Repository Software	CONTENTdm (local, hosted, pro)	12
Repository Software	Fedora	6
Repository Software	DSpace	6
Repository Software	ExLibris DigiTool	1
Repository Software	Innovative's Symposia	1
Repository Software	BePress Digital Commns	1
Repository Software	VITAL	1
	Repository Software Total	**28**
Image Repository	MDID	2
Image Repository	Luna	1
Image Repository	Artesia	1
Image Repository Total		4
Search Tools	Solr	1
Search Tools	DTSearch	1
	Search Tools Total	**2**
Delivery	EThOS (delivery)	1
Delivery	Bespoke Delivery	8
Delivery	Drupal	3
Delivery	Static Web pages	2
	Delivery Total	**14**
Journal Delivery	OJS	2
Journal Delivery Total		2
Preservation	Bespoke Preservation	4
Preservation	Quantum Digital Archive	1
	Preservation Total	**5**

Type of Software System	Specific System	Instances
3rd Party Delivery	JSTOR	2
3rd Party Delivery	Cengage	1
3rd Party Delivery	ProQuest	1
	3rd Party Delivery Total	**4**
A/V	iTunesU	2
A/V	Streaming Server	2
	A/V Total	**4**
File Server	File Server	17
	File Server Total	**17**
Catalogs	IRIS (MD in FMPro)	1
Catalogs	CALM	1
Catalogs	MODES Catalogue	2
Catalogs	Catalogue -- Unknown	2
Catalogs	Extensis Portfolio	2
Catalogs	Relational DB	1
Catalogs	Allegro DB	1
Catalogs	OPAC	1
Catalogs	Tec-Rec	1
Catalogs	SIFT	1
Catalogs	MINISIS	1
	Catalogs Total	**14**
	Grand Total	**94**

20. Appendix: Worksheet to Estimate Costs

Estimating Preservation Costs

Start-Up

	Title	Yearly Salary*	Weeks of Work	Cost
Staff	Developer			$ -
	System Administrator			$ -
	Collection Coordinator			$ -
	Manager			$ -
	Other			$ -
			Total Staff:	**$ -**

	Type	Cost
Hardware & Software	Server	
	Master Disks	
	Backup (e.g., disks)	
	Power	
	Network	
	Software Purchase (e.g., content management system, backup system)	
	Other	
	Total H/W and S/W:	**$ -**
	Total Start-Up:	**$ -**

Ongoing

	Title	Yearly Salary*	Weeks of Work	Yearly Cost
Staff	System Administrator			$ -
	Collection Coordinator			$ -
	Manager			$ -
	Other			$ -
			Total Staff:	**$ -**

	Type	Cost	Years to Replace	Yearly Cost
Hardware Replacement Costs	Server			
	Master Disks			
	Backup Disks			
	Other			
			Total Annual H/W Replacement Costs	**$ -**

	Type	Cost	Per Year Frequency	Yeary Cost
Annual Hardware & Software Costs	Backup Service Fee			$ -
	Power			$ -
	Network			$ -
	Software Licensing			$ -
	Hardware Licensing			$ -
	Other			$ -
			Total H/W & S/W Costs	**$ -**
			Annual Ongoing Costs	**$ -**

An Excel version of this worksheet is available for use on the Portico website.[53]

[53] http://www.portico.org/digital-preservation/wp-content/uploads/2011/03/NEH-White-Paper-on-Preserving-Digital-Content-Preservation-Costs-Worksheet.xlsx