Digital Preservation Case Studies: Preservation Activities at Portico

Sheila Morrissey
Senior Research Developer, Portico, ITHAKA

UN FAO
Digital Preservation and JHOVE2
Rome
May 24, 2011
“Digital Preservation is Everyone’s Problem …

BUT IT ISN’T THE SAME PROBLEM FOR EVERYONE!!”
ITHAKA is a not-for-profit organization that helps the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways.

We pursue this mission by providing innovative services that aid in the adoption of these technologies and that create lasting impact.
Ithaka S+R is a research and consulting service that focuses on the transformation of scholarship and teaching in an online environment, with the goal of identifying the critical issues facing our community and acting as a catalyst for change.

JSTOR is a research platform that enables discovery, access, and preservation of scholarly content.

Portico is a digital preservation service for e-journals, e-books, and other scholarly e-content.
Portico is among the largest community-supported digital archives in the world.

Working with libraries, publishers, and funders, we preserve e-journals, e-books, and other electronic scholarly content to ensure researchers and students will have access to it in the future.
An “Insurance Policy” for e-Content

Provide libraries with access to archived content when it becomes lost, orphaned or abandoned (regardless of libraries’ past or current subscription):

- Publisher ceases operation
- Publisher discontinues title
- Publisher drops back file

• Provide libraries with post-cancellation access – if publisher specifically names Portico
• About 90% of titles in Archive are covered by Portico post-cancellation access rights.
Delivery

Triggered Content

<table>
<thead>
<tr>
<th>Title</th>
<th>Trigger Date</th>
<th>Publisher</th>
<th>Holdings Available</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graft</td>
<td>2007/12</td>
<td>SAGE Publications</td>
<td>v. 4-6</td>
<td>2001-2003</td>
</tr>
<tr>
<td>Pain Reviews</td>
<td>2009/07</td>
<td>Hodder</td>
<td>v. 5-9</td>
<td>1998-2002</td>
</tr>
</tbody>
</table>

Post-Cancellation Access Requests

14 Titles with PCA

1 Institutions with PCA
Post-Cancellation Access

**E-Journals**
- 88% with PCA
- 12% without PCA

**E-Books**
- 87% with PCA
- 13% without PCA
Over 2,000 societies, and associations have committed content to Portico through 122 publishers agreements.

- E-journal titles: 12,142
- E-book titles: 73,298
- D-collections: 39
Portico Participating Publishers

- United States, 64
- United Kingdom, 24
- Australia, 5
- Germany, 5
- The Netherlands, 4
- Canada, 2
- New Zealand, 1
- Italy, 1
- Austria, 1
- United Arab Emirates, 1
- Sweden, 1
- Hungary, 1
- India, 1
- Egypt, 1
- Bangladesh, 1
- Other, 9

Numbers as of 8/31/2010
## Participating Libraries

<table>
<thead>
<tr>
<th>Participating Libraries</th>
<th>690</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Libraries</td>
<td>360</td>
</tr>
<tr>
<td>Non-US Libraries</td>
<td>330</td>
</tr>
</tbody>
</table>

Numbers as of 8/31/2010
Portico Participating Libraries

- United States, 360
- Brazil, 153
- Greece, 54
- Italy, 31
- Canada, 27
- United Kingdom, 22
- Australia, 20
- Ireland, 8
- New Zealand, 7
- Bangladesh, 1
- Cyprus, 1
- India, 1
- Lebanon, 1
- Sweden, 2
- Israel, 2
- Other, 8

Numbers as of 8/31/2010
TAKE THE LONG VIEW...
2002 Launch of Electronic Archiving Initiative by JSTOR

2005 Portico launched

2005 Portico signs initial e-journal publishers

2006 Portico ingests initial e-journal content into the archive

2008 Portico signs initial e-book publishers

2009 Portico ingests initial e-book content into the archive

2009 Portico signs initial d-collections

2010 Portico ingests initial d-collection content
Portico Participating Publishers

Numbers as of 8/31/2010
Portico Growth in Participating Titles

Participating E-Books

Participating E-Journals

Numbers as of 8/31/2010
Portico Participating Libraries

Numbers as of 10/29/2010
Types of Files Preserved

- Publisher Supplied Text: 27%
- Portico Created Archival Text: 25%
- Images: 48%
- Multi-file Packages: 0%
- Videos: 0%
- Audio: 0%
- Executable: 0%
- Application Specific Files: 0%
Mime Types Preserved

1. application/mathematica
2. application/msword
3. application/octet-stream
4. application/pdf
5. application/postscript
6. application/rtf
7. application/sgml
8. application/vnd.corel-presentations
9. application/vnd.ms-excel
10. application/vnd.ms-htmlhelp
11. application/vnd.ms-powerpoint
12. application/vnd.openxmlformats-
13. officedocument.wordprocessingm-
14. application/vnd.rn-realmedia
15. application/vnd.ms-excel
16. application/vnd.ms-htmlhelp
17. application/vnd.ms-powerpoint
18. application/vnd.openxmlformats-
19. officedocument.wordprocessingm-
20. application/x-asp
21. application/x-gzip
22. application/x-pc-cvs
23. application/x-pc-els
24. application/x-pc-eps
25. application/x-pc-gams
26. application/x-pc-msoffice
27. application/x-pc-netlogo
28. application/x-pc-nexus
29. application/x-pc-paintshoppro
30. application/x-pc-r
31. application/x-pc-stata
32. application/x-pc-stata-program
33. application/x-pc-tsp
34. application/x-pc-utf16
35. application/x-pc-utf8
36. application/x-rar-compressed
37. application/x-sgml-external-
parsed-entity
38. application/x-sh
39. application/x-shockwave-flash
40. application/x-tar
41. application/zip
42. audio/mpeg
43. audio/x-ms-wma
44. audio/x-wav
45. image/gif
46. image/jpeg
47. image/png
48. image/tiff
49. image/vnd.adobe.photoshop
50. image/x-ms-bmp
51. image/x-wmf
52. model/vrml
53. text/csv
54. text/html
55. text/plain
56. text/x-c++src
57. text/x-csrc
58. text/x-ptc-iso-8859
59. video/avi
60. video/mp4
61. video/mpeg
62. video/quicktime
63. video/x-flv
64. video/x-ms-wmv
# Preservation Levels on Files Preserved

<table>
<thead>
<tr>
<th>Preservation Level</th>
<th>Files</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>142,079,610</td>
<td>81.22%</td>
</tr>
<tr>
<td>Byte-Preserve</td>
<td>16,738,528</td>
<td>9.57%</td>
</tr>
<tr>
<td>System</td>
<td>14,869,679</td>
<td>8.50%</td>
</tr>
<tr>
<td>Reasonable-Effort</td>
<td>1,244,811</td>
<td>0.71%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>174,932,628</strong></td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
Preservation Levels on Files Preserved

- Full: 81%
- Byte-Preserve: 10%
- System: 8%
- Reasonable-Effort: 1%
<table>
<thead>
<tr>
<th>Format Status</th>
<th>Files</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well Formed and Valid</td>
<td>156,948,510</td>
<td>89.72%</td>
</tr>
<tr>
<td>Not Determined</td>
<td>16,304,477</td>
<td>9.32%</td>
</tr>
<tr>
<td>Well Formed and Not Valid</td>
<td>1,245,314</td>
<td>0.71%</td>
</tr>
<tr>
<td>Not Well Formed</td>
<td>434,074</td>
<td>0.25%</td>
</tr>
<tr>
<td>Well Formed</td>
<td>253</td>
<td>0.00%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>174,932,628</strong></td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
Format Status of Files Preserved

- Not Well Formed: 0%
- Well Formed and Not Valid: 1%
- Not Determined: 9%
- Well Formed and Valid: 90%
## Content Types of Files Preserved

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Files</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-journal Files</td>
<td>174,517,812</td>
<td>99.76%</td>
</tr>
<tr>
<td>Supplied E-journal Files</td>
<td>304,794</td>
<td>0.17%</td>
</tr>
<tr>
<td>E-book Files</td>
<td>108,829</td>
<td>0.06%</td>
</tr>
<tr>
<td>Technical Artifact Files</td>
<td>938</td>
<td>0.00%</td>
</tr>
<tr>
<td>Business Artifact Files</td>
<td>255</td>
<td>0.00%</td>
</tr>
<tr>
<td><strong>Total Files</strong></td>
<td><strong>174,932,628</strong></td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
Portico Technology Summary

OAIS-compliant repository designed for managed preservation

Key influences:
» OAIS, GDFR, PRONOM, PREMIS, METS, DC, NLM (JATS), MPEG-21 DIDL, ARK

Key technologies:
» XML, XML schema, Schematron, JHOVE, NOID
» Documentum, Oracle, Java, JMS, LDAP
» Format Registry
Portico Technology Summary

Archive design goals:
» Content preserved in application-neutral form using open standards
  • METS, PREMIS, JHOVE
» A “Bootstrapable Archive”: XML plus Digital Objects
  • Cached in Documentum and Oracle; replicated on file systems

Ingest system design goals:
» Pluggable tools to facilitate new providers and replacement tools
» Configurable workflows for different content types
» Scalable to very high content volumes
» Built on Documentum workflows
**Access Options**

- Audit
- Migrate
- Validate
- Fixity Check
- Completeness Check
- Repair
- Migrate/Normalize
- Track Events
- Ingest

**Preservation Actions**

- ConPrep System
- Archive Management System
- Content ingest & normalization
- Content preservation

**Content Receipt**

- Publisher
- Content packaging & delivery

**Access Options**

- Access
- Provided by JSTOR

**Preservation Planning**

- Study
- Monitor
- Plan
- Policy Definition
- Document
- Engage Community

**Data Flow & Systems**

- Content ingest & normalization
- ConPrep System
- Archive Management System

**Portico & Managed Preservation**

- Publisher
- Content packaging & delivery
- ConPrep System
- Archive Management System
- Access
- Provided by JSTOR
Publisher supplies XML Source file (including the text, images) and PDF page rendition.

Best approach for preserving the intellectual content of the article or book.

Authenticate: verify that preserved content is what it purports to be.

- Verify format: ensure the file meets syntactic and semantic rules of format specification.
- Repair
- Normalize (XML)
- Create preservation metadata

Assess archival robustness of file format.

- Migrate files to ensure future usability of content.
- Replicate objects and metadata to protect against bit rot and media deterioration

Render articles to meet viewing requirements of delivery platform.

Assess archival robustness of file format.
Portico E-Journal/E-Book Preservation Process

» Interviews with publisher production and technology staff
  • Formats used, production process, content delivered
  • Number of different types of content
  • Updates
  • Supplemental files

» Large sample data evaluation

» Formal (written) preservation action plan for each publisher

» Tool development (as needed per preservation plan)

» Extensive automated QC during ingest
Portico Systems Overview

Content Providers (Publishers) → Content Setup System → Content Ingest System → Archive Management System → Delivery System (JSTOR) → Content Consumers (Universities, Scholars)

Sample content → Content to be archived → New Tools, Workflows, Configuration Data

New Tools, Workflows, Configuration Data

Content Providers

Sample content

Content to be archived

New Tools, Workflows, Configuration Data
The Problem: Publisher content with arbitrary naming conventions and packaging rules
The Solution: Profiles containing publisher-specific policies, defaults, and overrides
The Implementation: XML instances, java regular expression patterns to tokenize complex file names

Design and Coding by Roland Mesde
Digital Preservation is Everyone’s Problem …

MEANS: “YOU ARE NOT ALONE!!”
Standards and Community Activities

» NLM DTD Advisory Board
» NISO Standards Architecture Committee
» NISO Journal Article Versions Working Group (completed)
» PREMIS Working Group (completed)
» Global Digital Format Registry (now UDFR)
» PEPRS (Piloting an e-journals preservation registry service)
» DPC (Digital Preservation Coalition)
» NDSA (National Digital Stewardship Alliance)
Portico Research and Community Activities

Grant-Funded Projects

» NDIIPP Grant to Portico

» JISC Digitisation Programme Preservation Study
  • Univ. of London Computer Centre, Portico, and Digital Preservation Coalition

» IMLS Project on digital book preservation
  • Cornell Univ. and Portico

» JHOVE2 project (NDIIPP-funded)
  • California Digital Library, Portico, Stanford Digital Repository
Internal Projects

» E-Book study
» Library-created content study
» Portico Preservation Metadata 2.0
The Center for Research Libraries (CRL) conducted a preservation audit of Portico (www.portico.org) between April and October 2009 and hereby certifies Portico as a trustworthy digital repository. CRL has found that Portico’s services and operations basically conform to the requirements for a trusted digital repository. The CRL Certification Advisory Panel has concluded that the practices and services described in Portico’s public communications and published documentation are generally sound and appropriate to the content being archived and the needs of the CRL community. Moreover the CRL Certification Advisory Panel expects that in the future, Portico will continue to be able to deliver content that is understandable and usable by its designated user community.

This certification is based upon a site visit and sampling of archives content, and upon the review of information gathered by CRL and its Certification Advisory Panel and of documents and documentation provided by Portico. CRL’s analysis was guided by the criteria included in the Trustworthy Repositories Audit and Certification checklist, and other metrics developed by CRL through its analyses of digital repositories.

CRL conducted its audit with reference to generally accepted best practices in the management of digital systems; and with reference to the interests of its community of research libraries and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada. The purpose of the audit was to obtain reasonable assurance that Portico provides, and is likely to continue to provide, services adequate to those needs without material flaws or defects and as described in Portico’s public disclosures. The CRL audit provides a reasonable basis for these findings.

CRL has assigned Portico the following levels of certification (the numeric rating is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level):

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>PORTICO SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational Infrastructure</td>
<td>3</td>
</tr>
<tr>
<td>Digital Object Management</td>
<td>4</td>
</tr>
<tr>
<td>Technologies, Technical Infrastructure, Security</td>
<td>4</td>
</tr>
</tbody>
</table>
Life is messy...

... FOR EVERYONE!!
Non-standard packaging

Incoming File System

PublisherA

0008543x
2006
106

B

CNCR21779
21779_ftp.pdf
21779_ftp.sgm

equation

aueq001.tif
aueq002.tif
nueq001.gif
nueq002.gif

image_m
mfig001.jpg

image_n
nfig001.jpg

image_t
tfig001.gif

Resulting Content Model

Content Unit (Article)

Text: Marked Up Text

21779_ftp.sgm

Rendition: Page Images

21779_ftp.pdf

Component: Formula Graphic

aueq001.tif
nueq001.gif

Component: Formula Graphic

aueq002.tif
nueq002.gif

Component: Figure Graphic

mfig001.jpg
nfig001.jpg
tfig001.gif
Content isn’t perfect
  » Must have policies and workflow for invalid data
  » There are degrees of “badness”
  » Strict format validity does not equate to usefulness or usability
    • E.g., Well-formed but not valid PDF
    • E.g., Valid PDF with bad embedded font
    • E.g., Invalid JPEG

Content creation practices change over time
  » Publishers (content providers) aren’t consistent
  » Or don’t warn you that they are changing something
  » Defensive programming required

Software isn’t perfect
  » Assume that there will be internal failures
  » Reversibility and audit trail are essential
Moving to Preservation at Scale

» Scale up from 900K articles/year to 10 million articles/year

» Involved changes to
  • Software
  • Hardware
  • Procedures

» Testing, tuning
  • How many threads?
  • Good data, bad data
  • More batches? Bigger batches?
  • Long-running tests

» Side effects
  • Loaders
  • Cleanup
  • Logging
  • User interface
  • Storage backup and recovery
Monthly Article Ingest versus Capacity
Some links

» Portico Web Site
  • Portico TRAC self-audit
  • Portico Policy Documents

» CRL Audit Report

» NEH Report on Preservation of Books and Other Digital Content

» Blue Ribbon Task Force on Sustainable Digital Preservation and Access

» NLM Journal Archiving and Interchange Tag Suite
THANK YOU.

Sheila Morrissey
sheila.morrissey@ithaka.org

http://www.portico.org