I T H A K A

JSTOR | PORTICO | ITHAKA S+R

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!!”

PORTICO

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

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.

PORTICO

Delivery

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Triggered Content

Title	Trigger Date	Publisher	Holdings Available	Years
Auto/Biography	2008/07	SAGE Publications	v. 12-14	2004-2006
Brief Treatment and Crisis Intervention	2009/04	Press	v. 1-8	2001-2008
Graft	2007/12	SAGE Publications	v. 4-6	2001-2003
Pain Reviews	2009/07	Hodder	v. 5-9	1998-2002

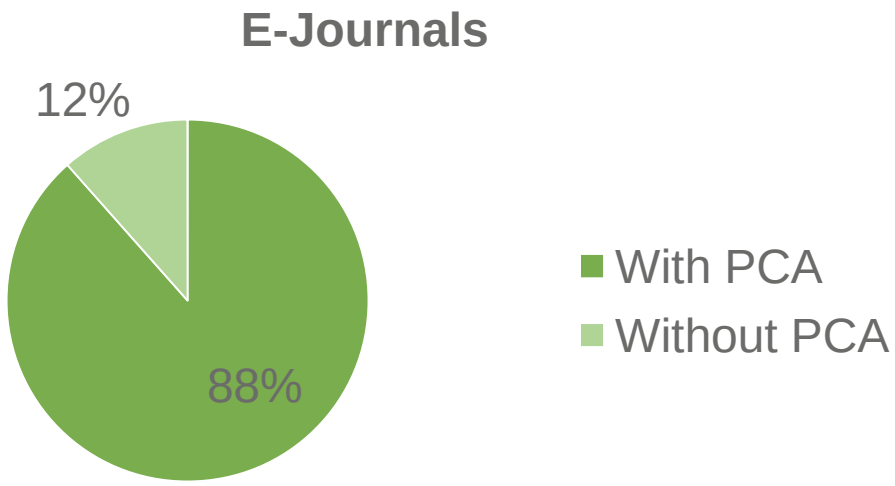
Post-Cancellation Access Requests

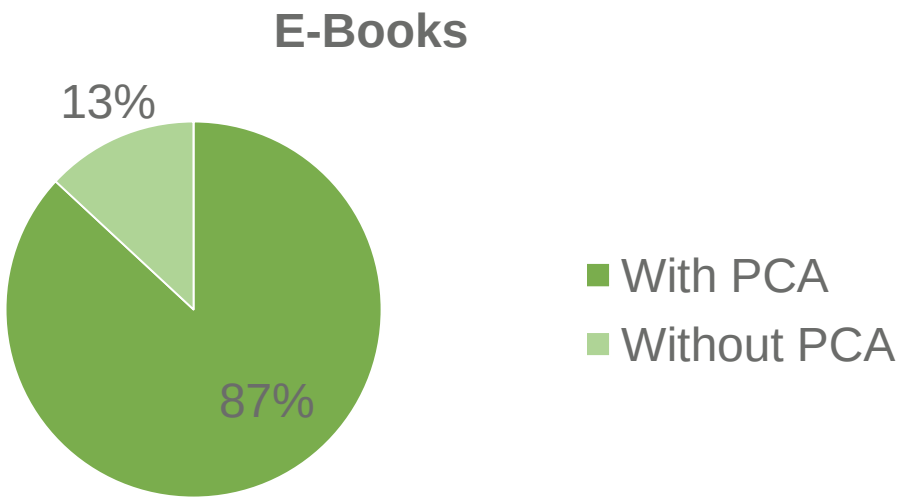
14 Titles with PCA

1 Institutions with PCA

PORTICO

Post-Cancellation Access



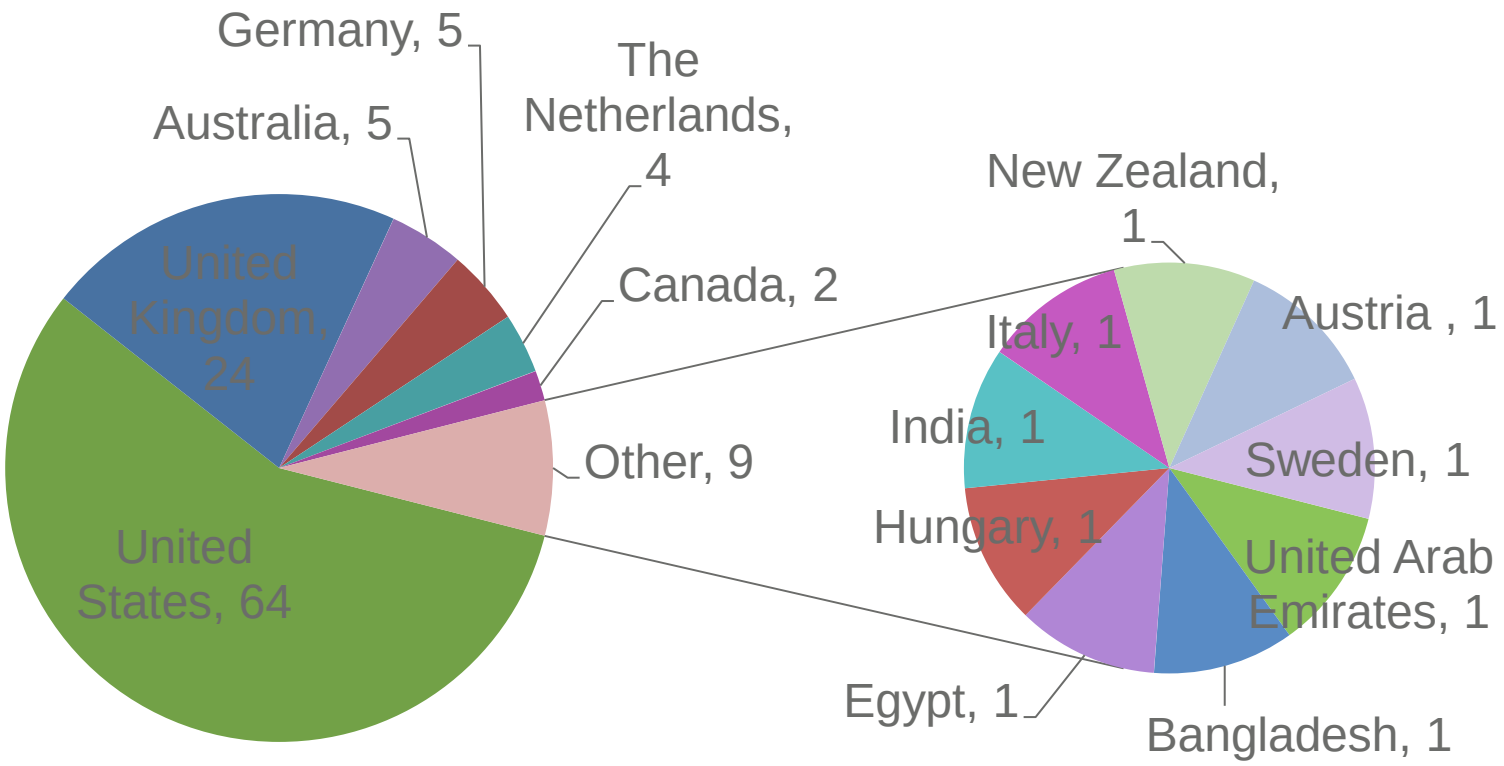


PORTICO

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



Numbers as of 8/31/2010

PORTICO

Participating Libraries

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

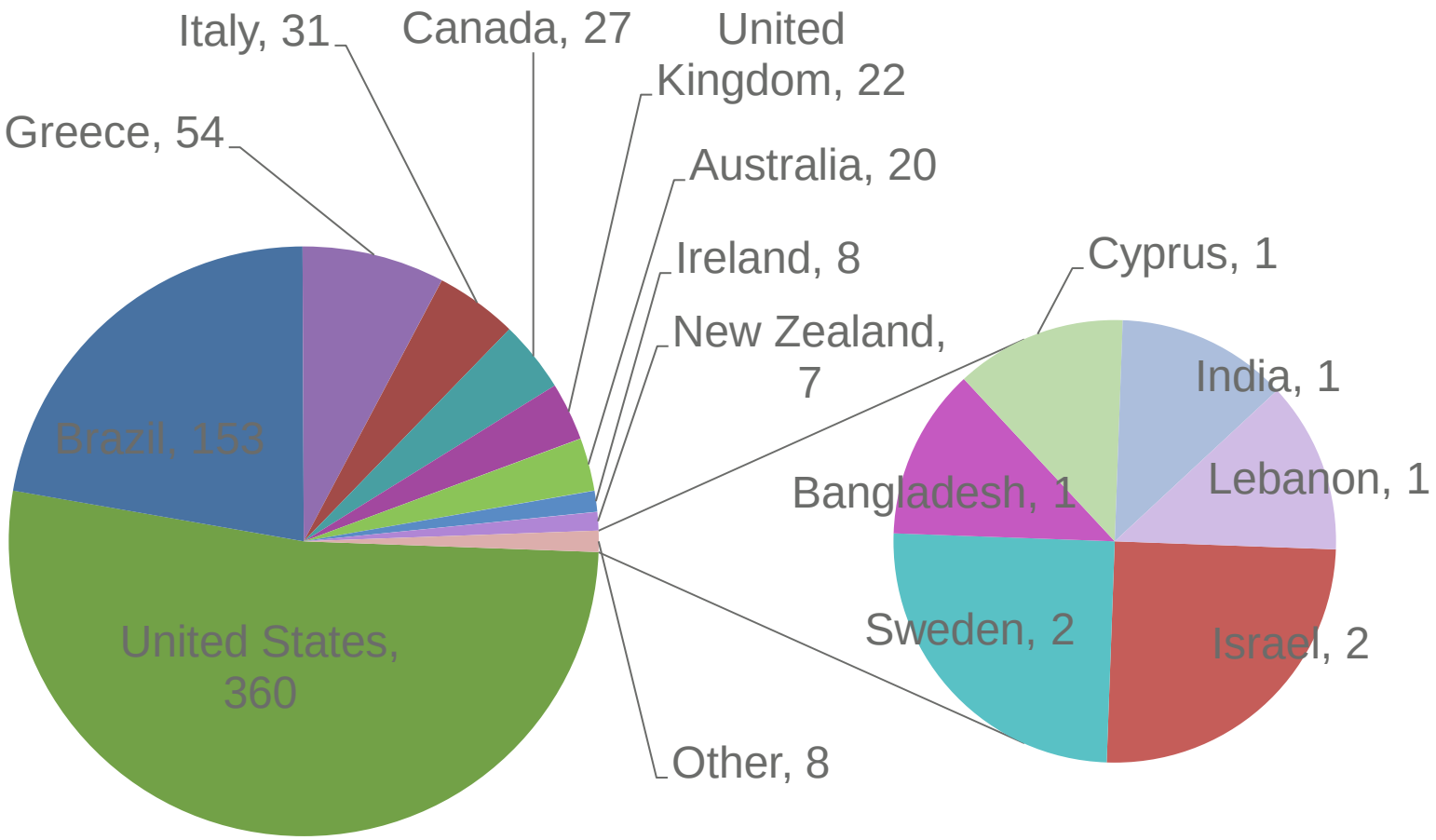
Participating Libraries	
Participating Libraries	690
US Libraries	360
Non-US Libraries	330

Numbers as of 8/31/2010

PORTICO

Portico Participating Libraries

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



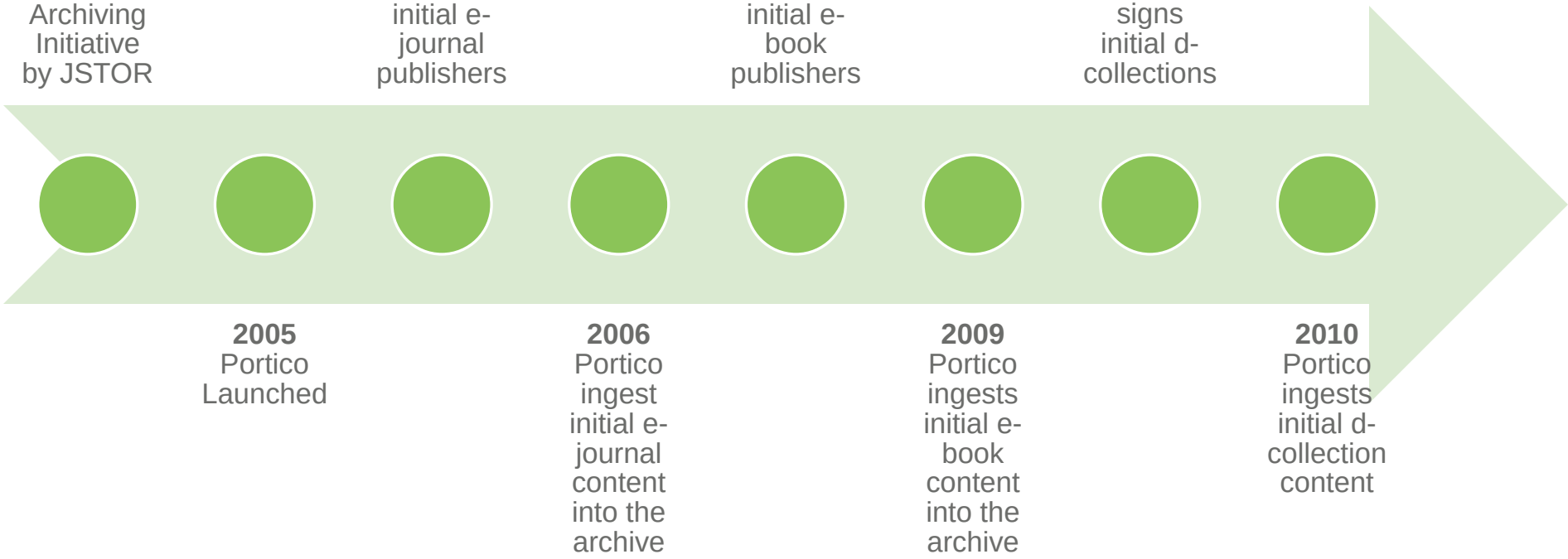
Numbers as of 8/31/2010

PORTICO

TAKE THE LONG VIEW…

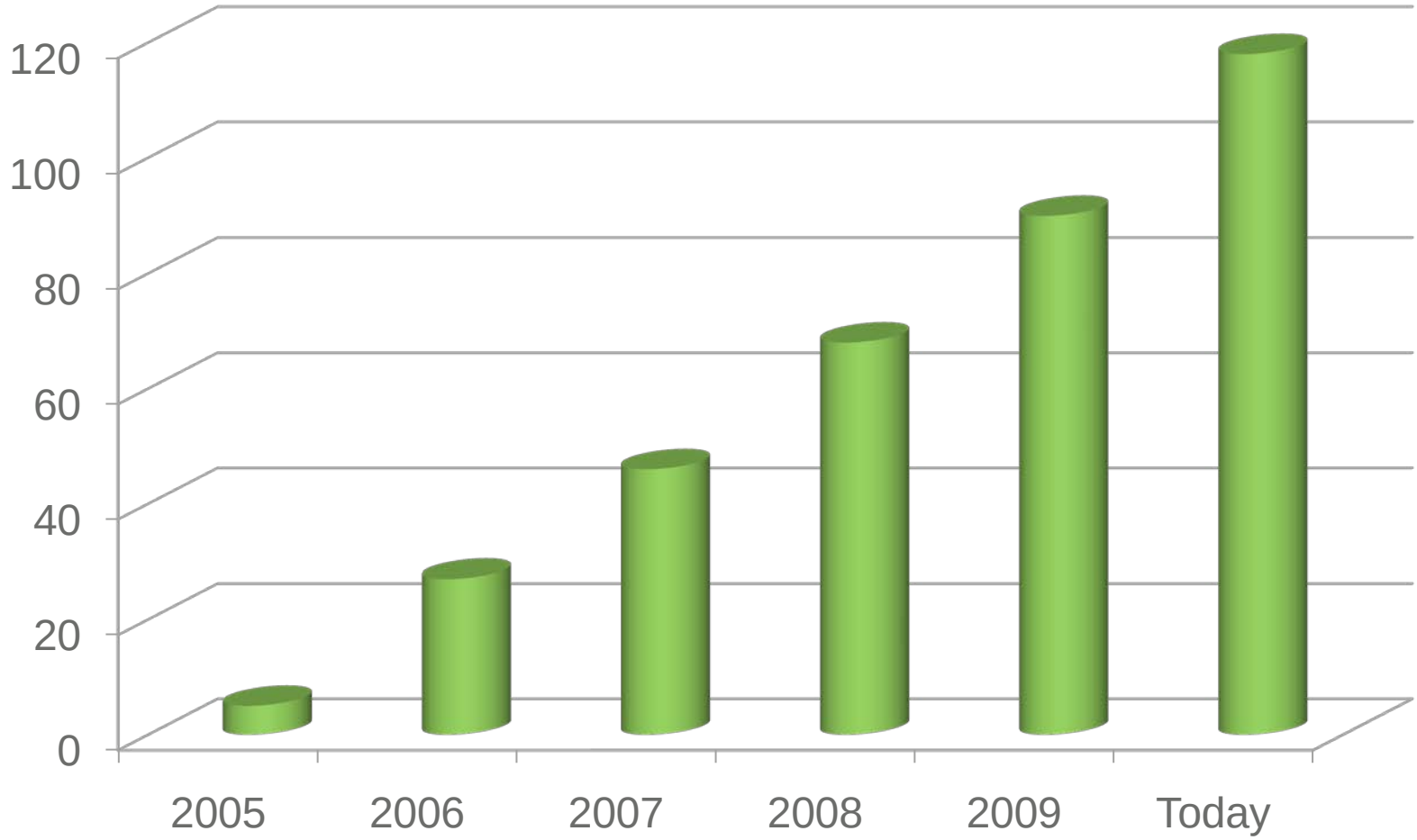
Portico Timeline

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



PORTICO

Portico Participating Publishers

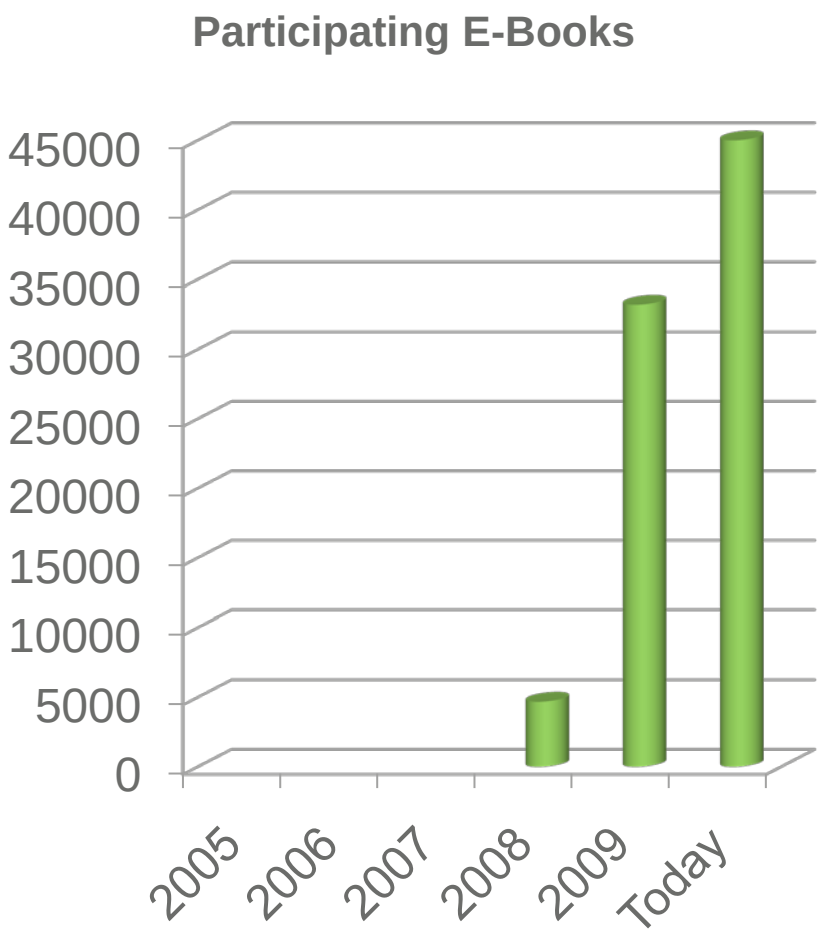


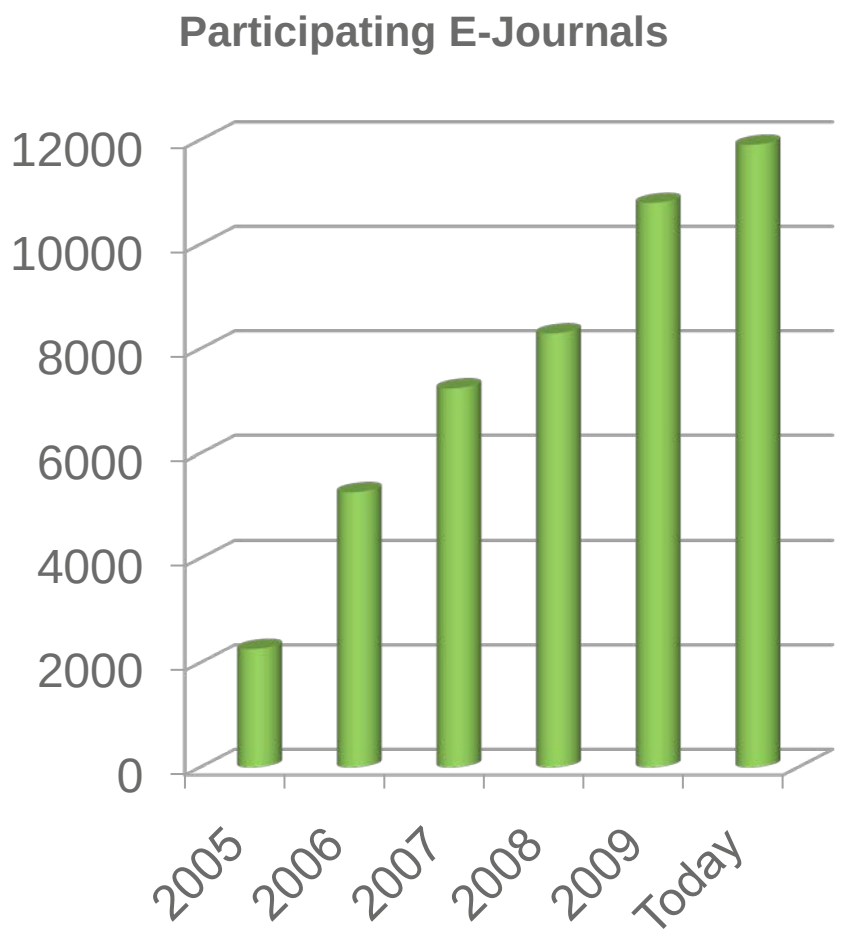
Numbers as of 8/31/2010

PORTICO

Portico Growth in Participating Titles

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



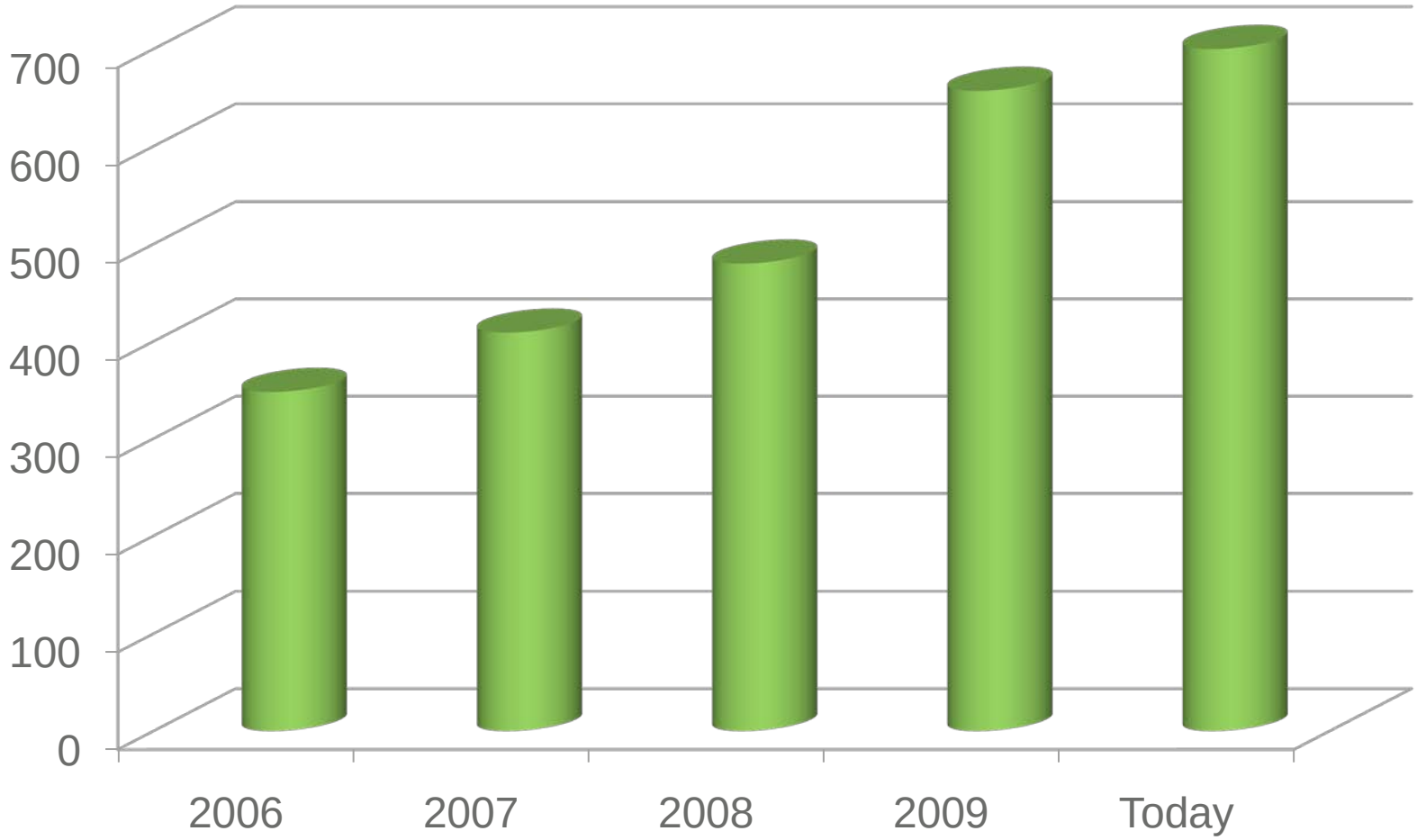


Numbers as of 8/31/2010

PORTICO

Portico Participating Libraries

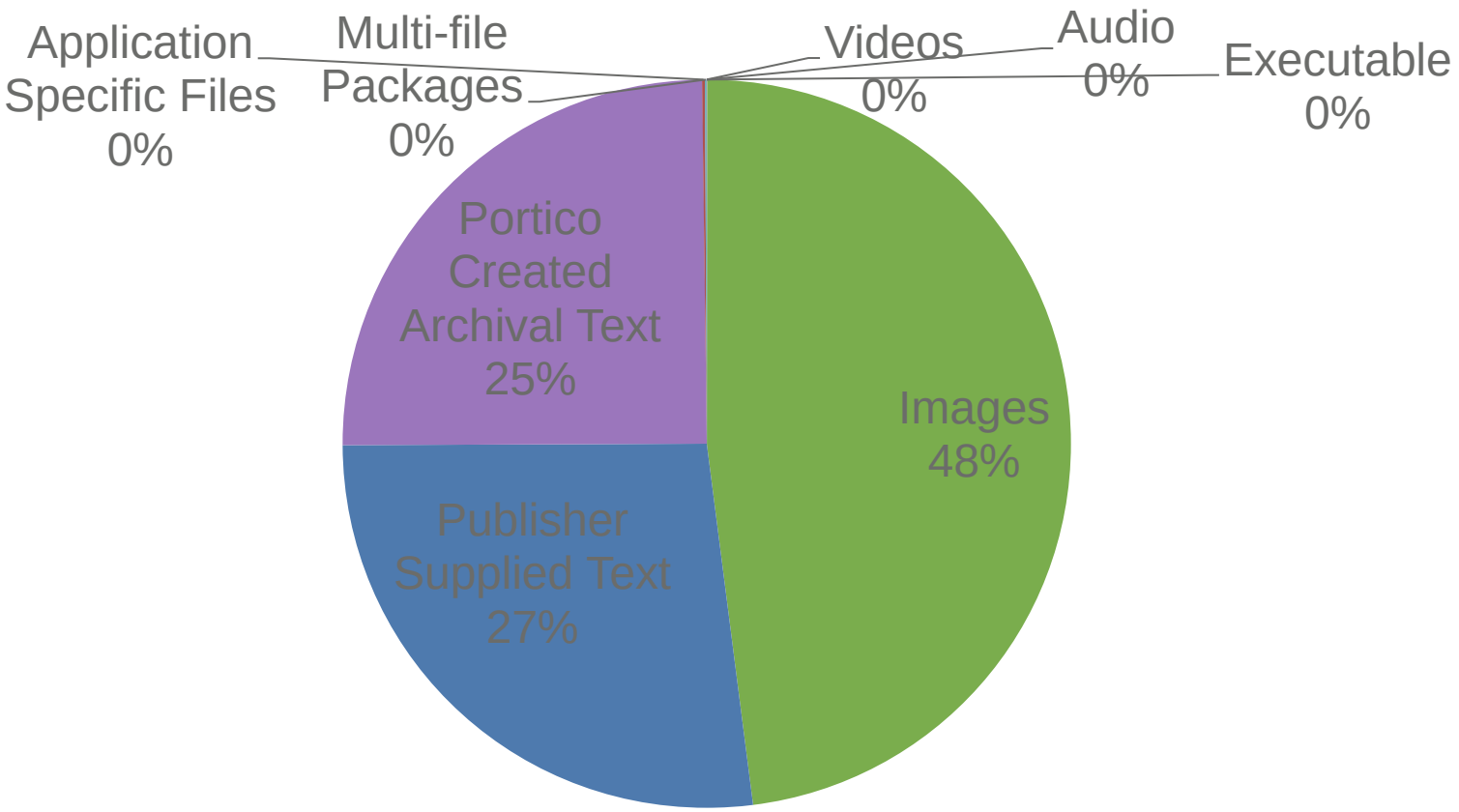
ITHAKA
JSTOR | PORTICO | ITHAKA S+R



Numbers as of 10/29/2010

PORTICO

Types of Files Preserved



Mime Types Preserved

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

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.wordprocessingml.document
14. application/vnd.rn-realmedia
15. application/vnd.wordperfect
16. application/x-asp
17. application/x-gzip
18. application/x-mathcad
19. application/xml
20. application/xml-dtd
21. application/xml-external-parsed-entity
22. application/x-ptc-els-Application Specific Filesset-toc-snippet
23. application/x-ptc-els-Application Specific Filesset-toc-xml-snippet
24. application/x-ptc-eps
    application/x-ptc-exe
25. application/x-ptc-gams
26. application/x-ptc-msoffice
27. application/x-ptc-netlogo
28. application/x-ptc-nexus
29. application/x-ptc-paintshoppro
30. application/x-ptc-r
31. application/x-ptc-stata-Application Specific Files
32. application/x-ptc-stata-program
33. application/x-ptc-tsp
34. application/x-ptc-utf16
35. application/x-ptc-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

PORTICO

Preservation Levels on Files Preserved

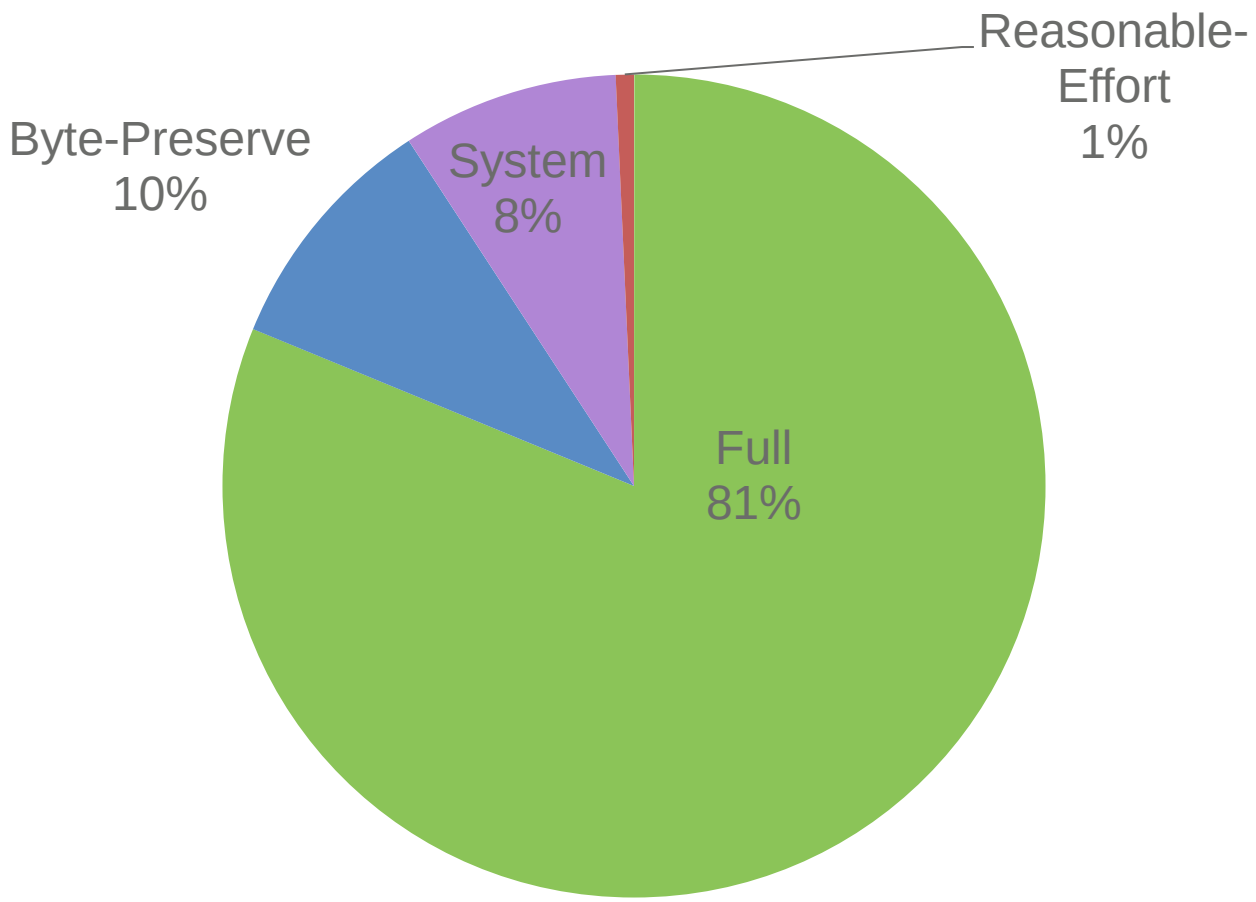
ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Preservation Level	Files	Percent
Full	142,079,610	81.22%
Byte-Preserve	16,738,528	9.57%
System	14,869,679	8.50%
Reasonable-Effort	1,244,811	0.71%
Total	**174,932,628**	**100.00%**

PORTICO

Preservation Levels on Files Preserved

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



PORTICO

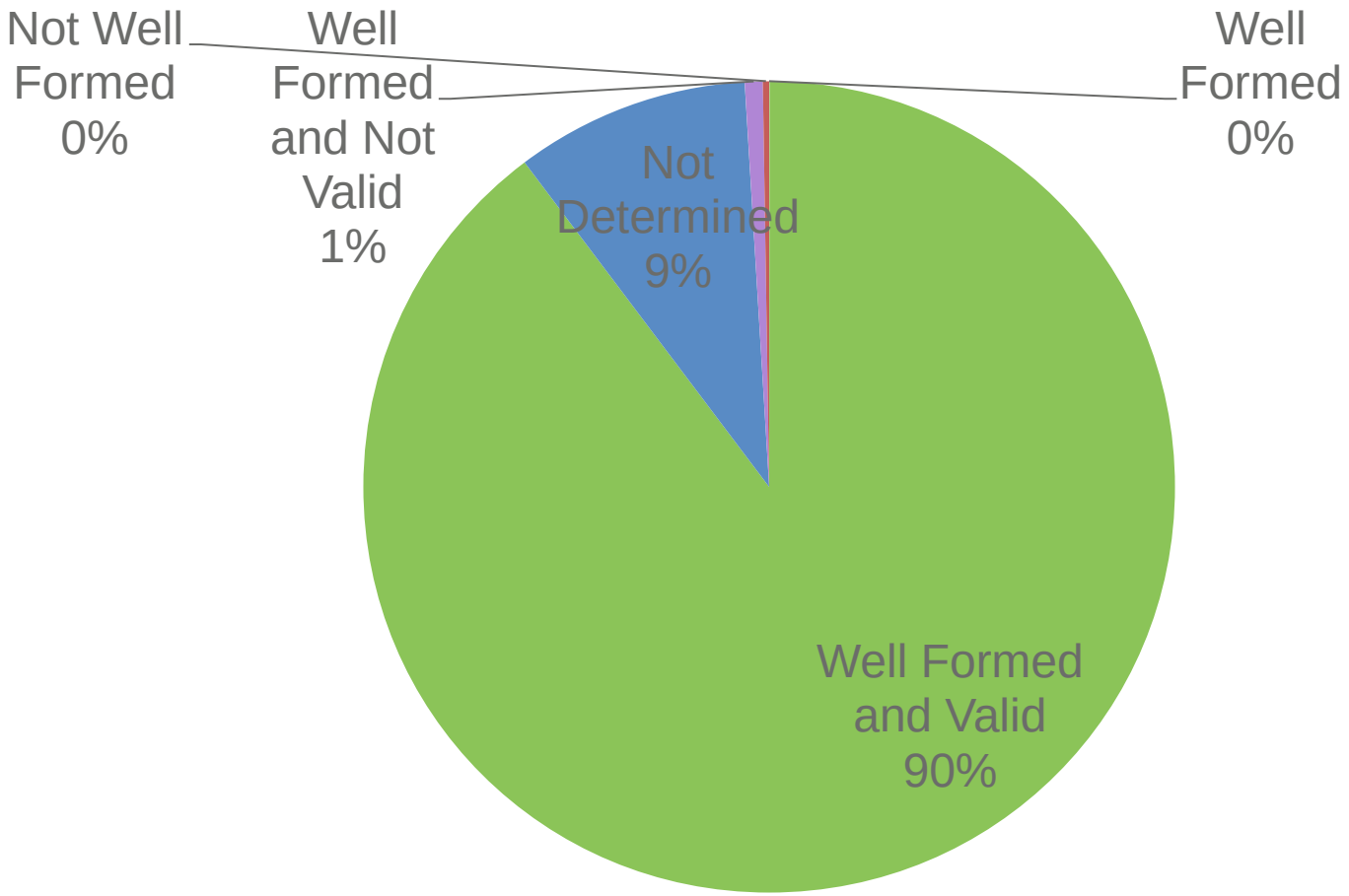
Format Status of Files Preserved

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Format Status	Files	Percent
Well Formed and Valid	156,948,510	89.72%
Not Determined	16,304,477	9.32%
Well Formed and Not Valid	1,245,314	0.71%
Not Well Formed	434,074	0.25%
Well Formed	253	0.00%
Total	**174,932,628**	**100.00%**

PORTICO

Format Status of Files Preserved



Content Types of Files Preserved

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Content Type	Files	Percent
E-journal Files	174,517,812	99.76%
Supplied E-journal Files	304,794	0.17%
E-book Files	108,829	0.06%
Technical Artifact Files	938	0.00%
Business Artifact Files	255	0.00%
Total Files	**174,932,628**	**100.00%**

PORTICO

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

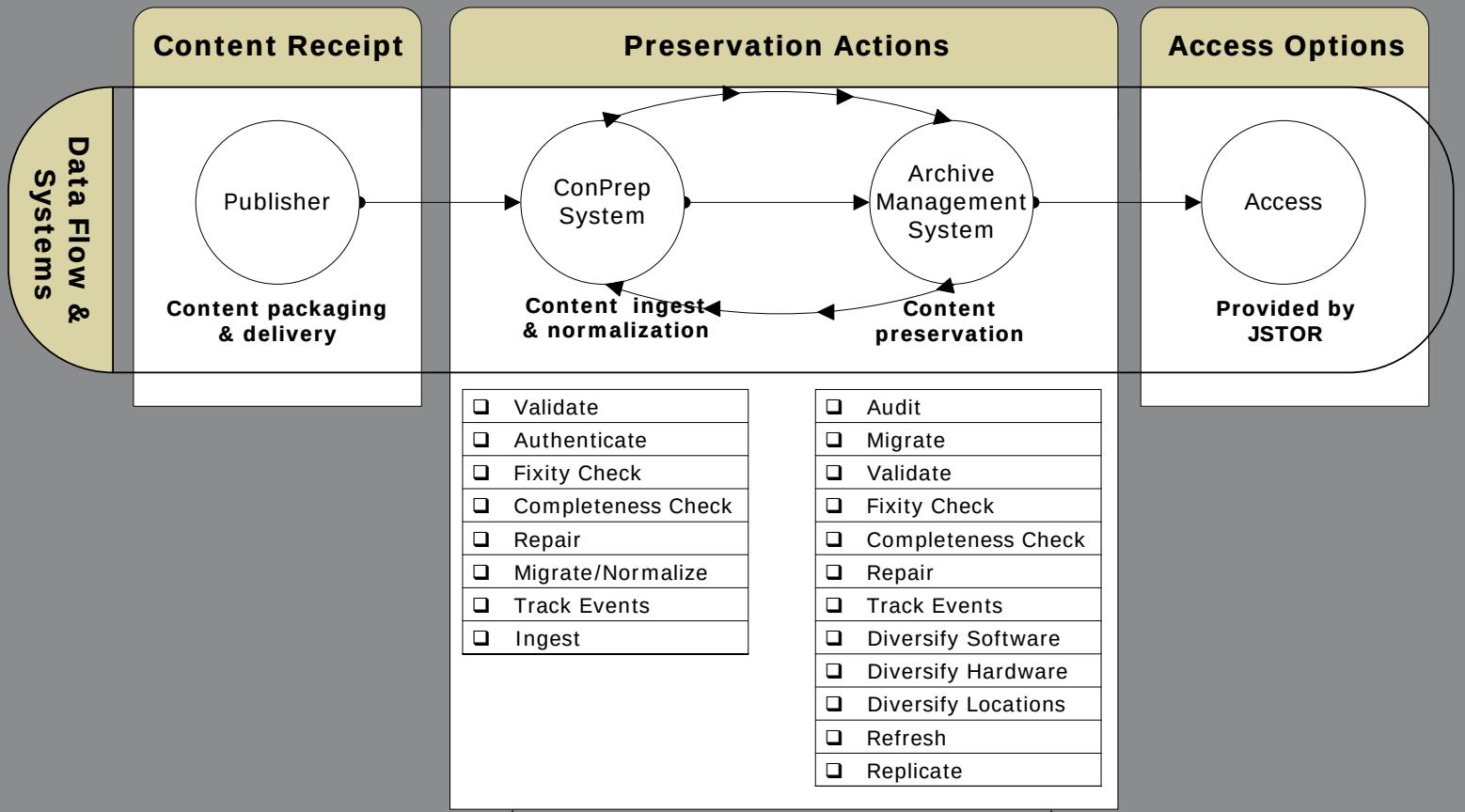
Portico Technology Summary

Archive design goals:

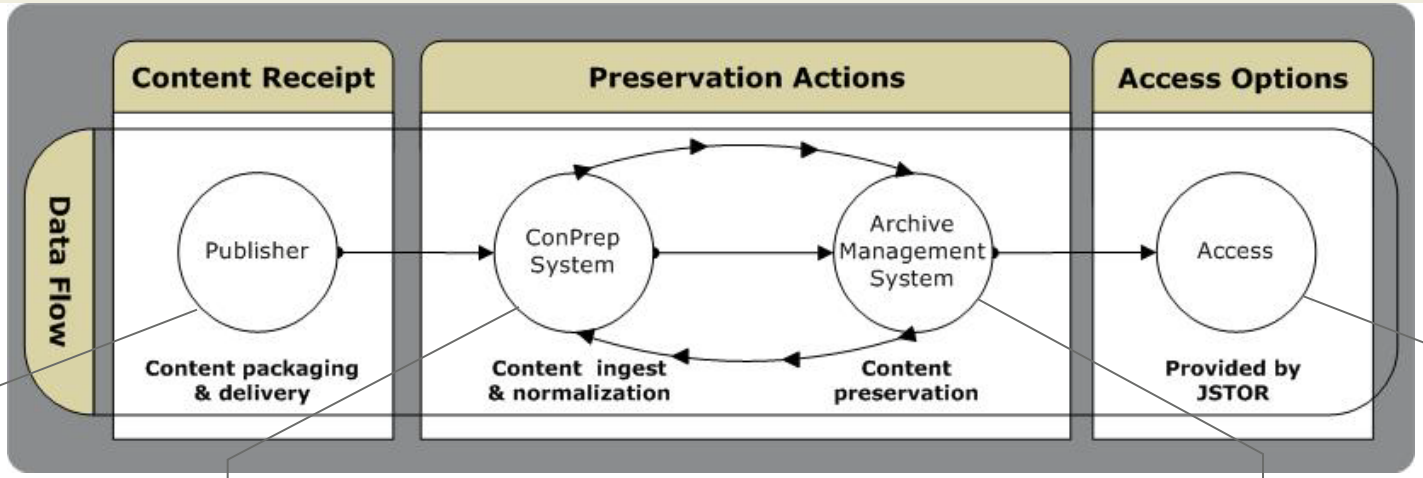
- Content preserved in application-neutral form using open standards
  - METS, PREMIS, JHOVE
- A “Bootstrapable Archive”: XML plus Digital Objects
  - Cached in Documentum and Oracle; replicated on file systems

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



Portico & Managed Preservation



- 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

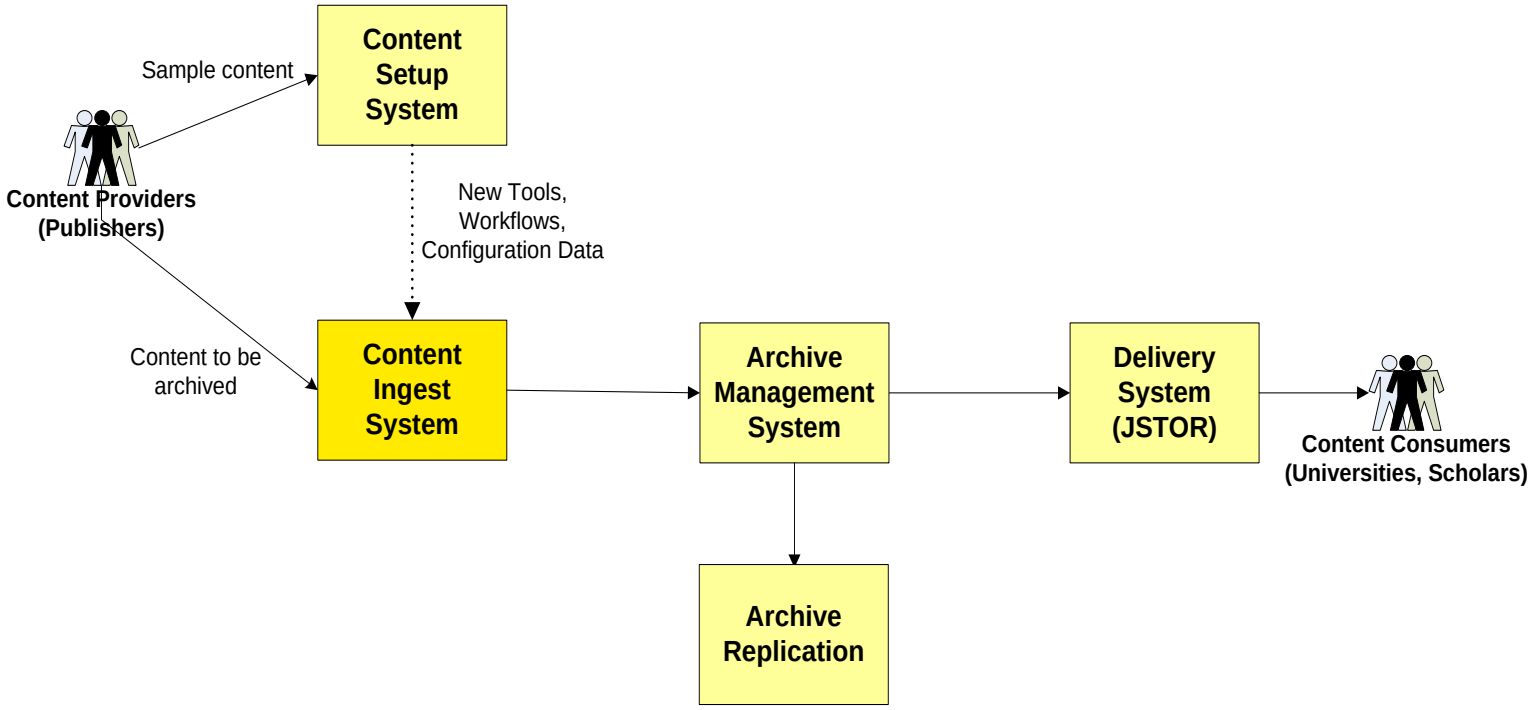
- Render articles to meet viewing requirements of delivery platform.

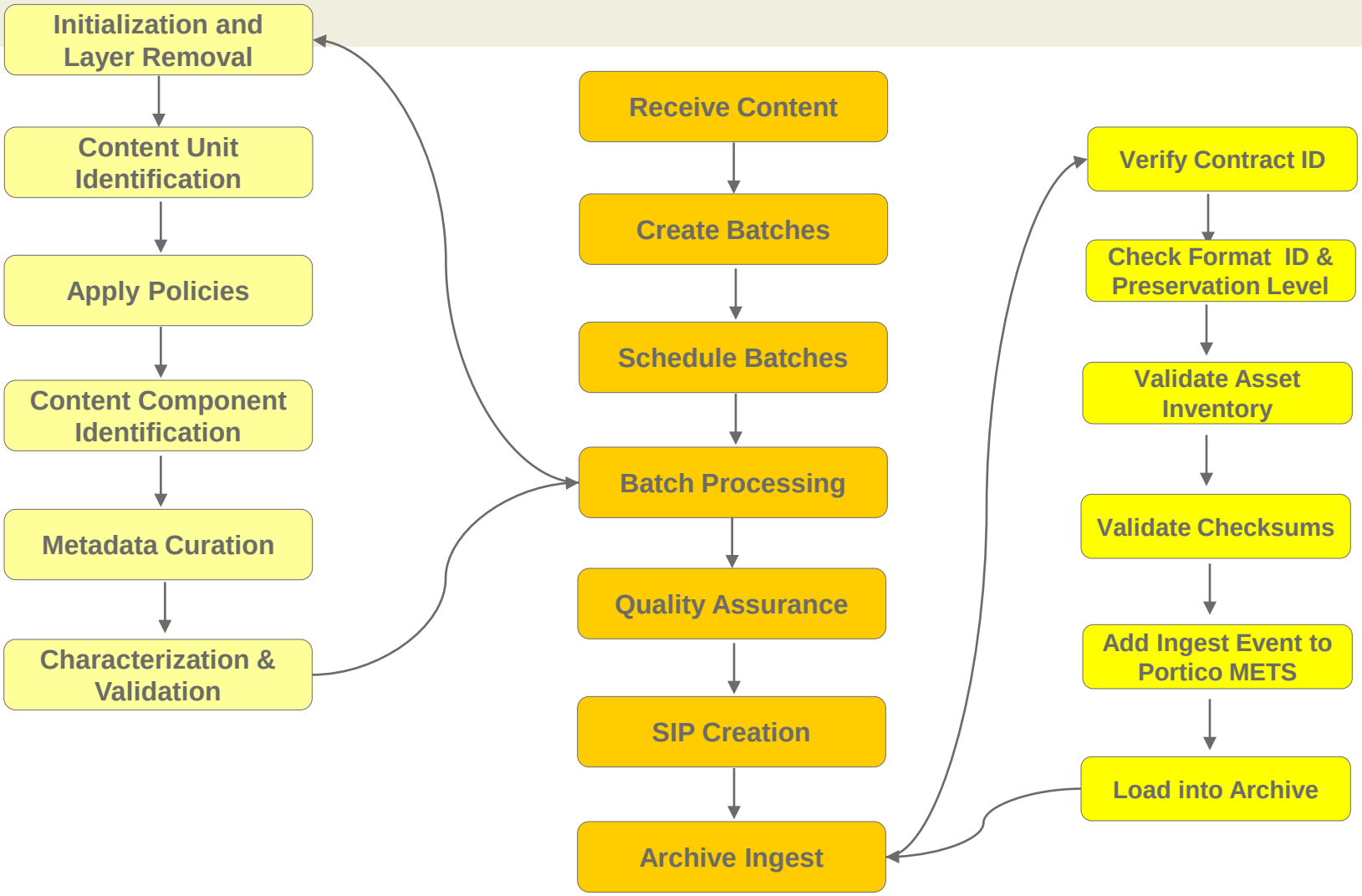
PORTICO

Portico E-Journal/E-Book Preservation Process

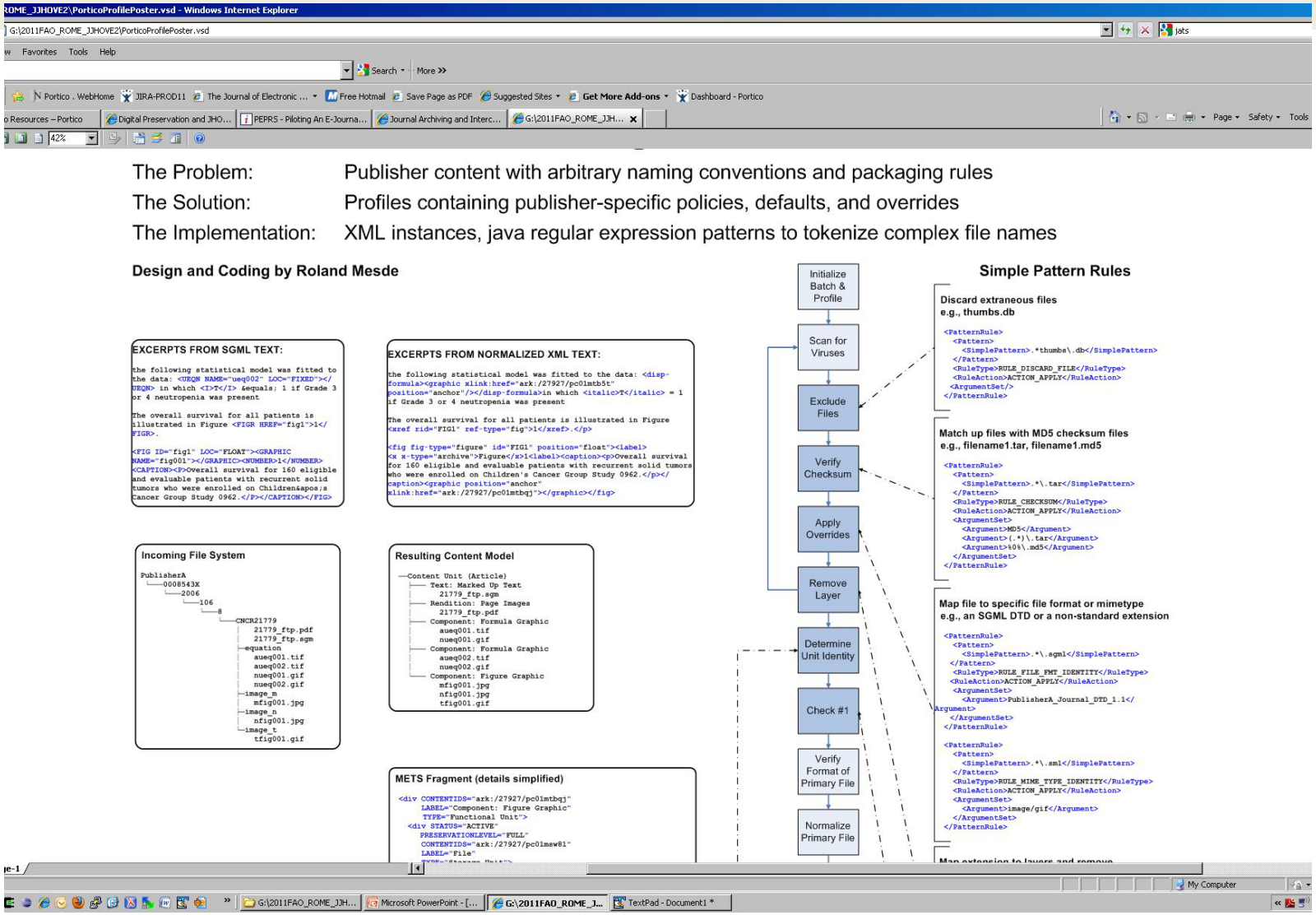
- Interviews with publisher production and technology staff
  - Formats used, production process, content delivered
  - Number of different types of content
  - Updates
  - Supplemental files
- Large sample data evaluation
- Formal (written) preservation action plan for each publisher
- Tool development (as needed per preservation plan)
- Extensive automated QC during ingest

Portico Systems Overview




Initialization and Layer Removal
Content Unit Identification
Apply Policies
Content Component Identification
Metadata Curation
Characterization & Validation
Receive Content
Create Batches
Schedule Batches
Batch Processing
Quality Assurance
SIP Creation
Archive Ingest
Verify Contract ID
Check Format ID & Preservation Level
Validate Asset Inventory
Validate Checksums
Add Ingest Event to Portico METS
Load into Archive

PORTICO


ROME_JJHOVE2\PorticoProfilePoster.vsd - Windows Internet Explorer
G:\2011FAO_ROME_JJHOVE2\PorticoProfilePoster.vsd
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
EXCERPTS FROM SGML TEXT:
EXCERPTS FROM NORMALIZED XML TEXT:
Incoming File System
Resulting Content Model
METS Fragment (details simplified)
Initialize Batch & Profile
Scan for Viruses
Exclude Files
Verify Checksum
Apply Overrides
Remove Layer
Determine Unit Identity
Check #1
Verify Format of Primary File
Normalize Primary File
Simple Pattern Rules
Discard extraneous files
e.g., thumbs.db
Match up files with MD5 checksum files
e.g., filename1.tar, filename1.md5
Map file to specific file format or mimetype
e.g., an SGML DTD or a non-standard extension

Digital Preservation is Everyone's Problem …

MEANS: "YOU ARE NOT ALONE!!"

Portico Research and Community Activities

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

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

Portico Research and Community Activities

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

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

PORTICO

Portico Research and Community Activities

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

Internal Projects

- E-Book study
- Library-created content study
- Portico Preservation Metadata 2.0

PORTICO

CRL TRAC Audit

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

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):[1]

CATEGORY	PORTICO SCORE
Organizational Infrastructure	3
Digital Object Management	4
Technologies, Technical Infrastructure, Security	4

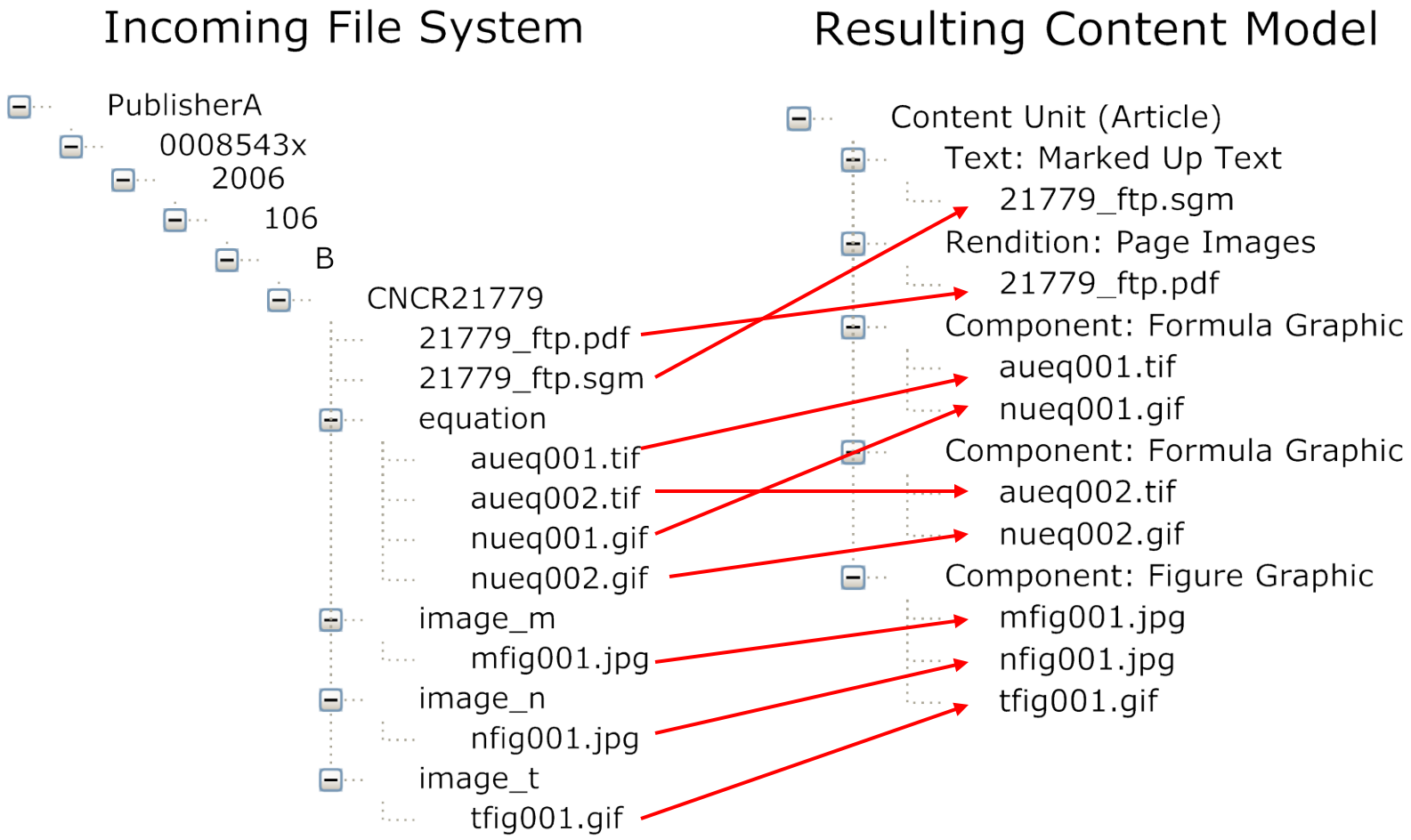
PORTICO

Life is messy…

… FOR EVERYONE!!

Non-standard packaging

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



PORTICO

“This is reality, Brad”

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

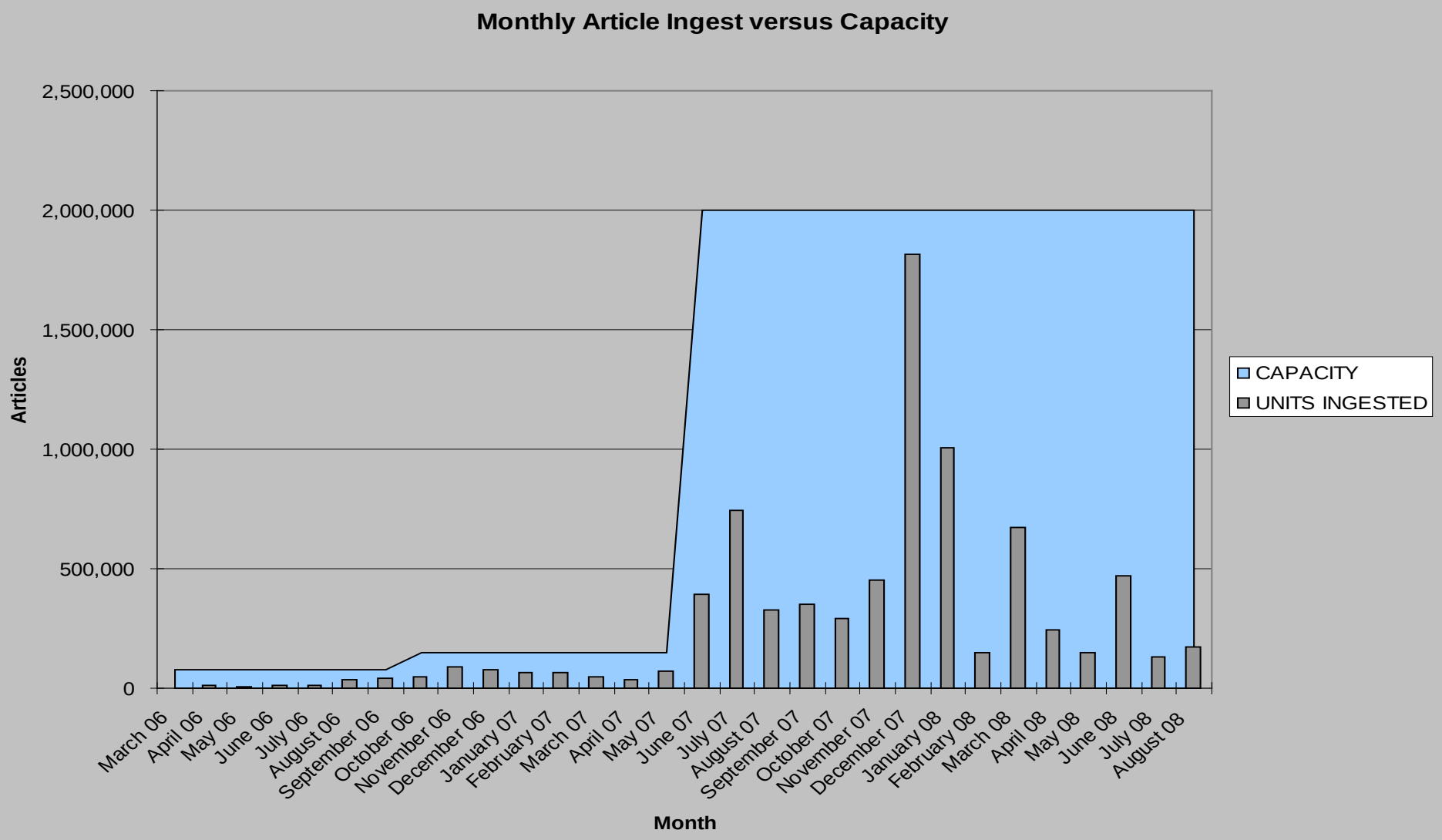
- Assume that there will be internal failures
- Reversibility and audit trail are essential

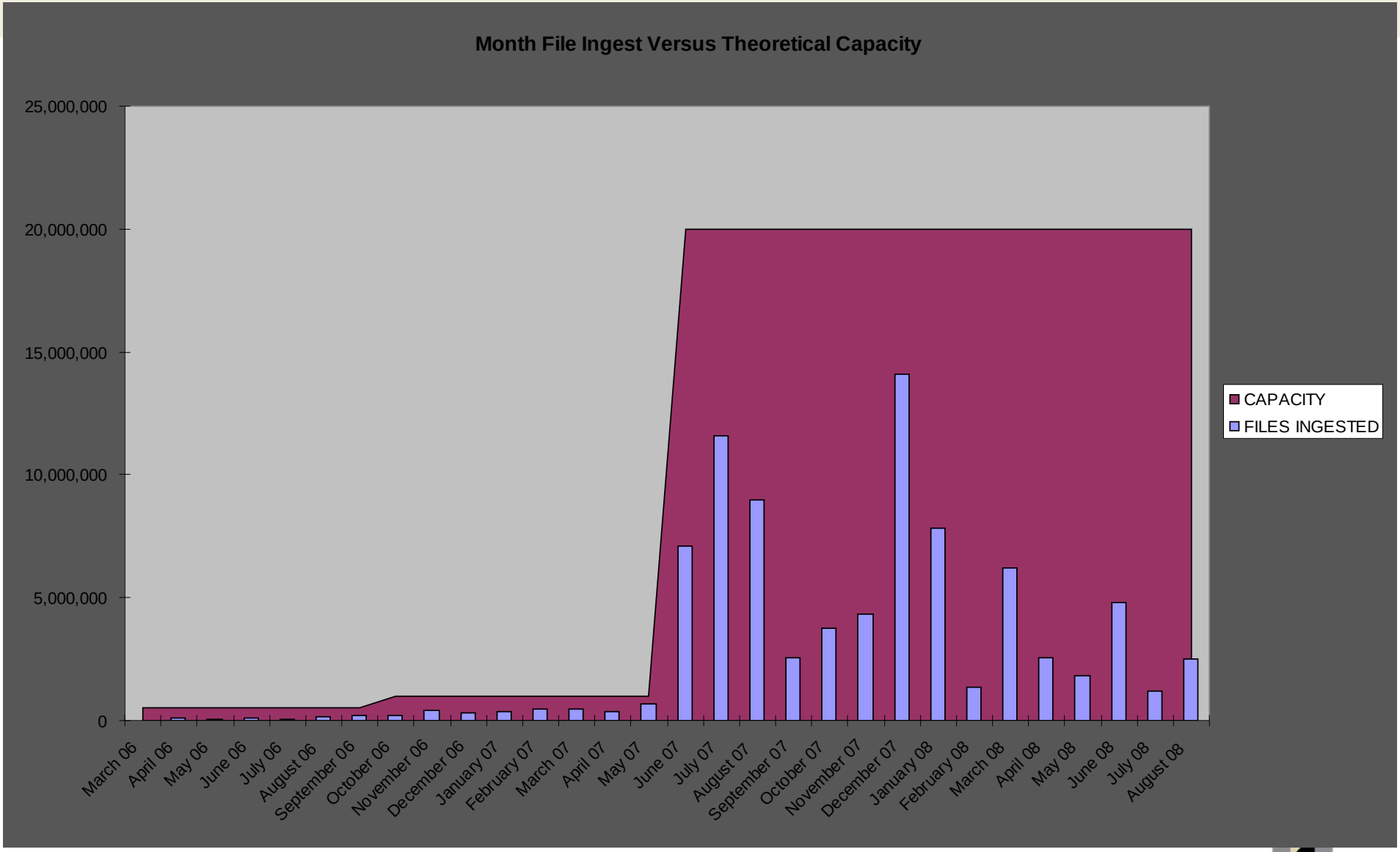
Moving to Preservation at Scale

ITHAKA
JSTOR | PORTICO | ITHAKA S+R

- 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

PORTICO


Monthly Article Ingest versus Capacity
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Articles
CAPACITY
UNITS INGESTED
March 06
April 06
May 06
June 06
July 06
August 06
September 06
October 06
November 06
December 06
January 07
February 07
March 07
April 07
May 07
June 07
July 07
August 07
September 07
October 07
November 07
December 07
January 08
February 08
March 08
April 08
May 08
June 08
July 08
August 08
Month


Month File Ingest Versus Theoretical Capacity
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
CAPACITY
FILES INGESTED
March 06
April 06
May 06
June 06
July 06
August 06
September 06
October 06
November 06
December 06
January 07
February 07
March 07
April 07
May 07
June 07
July 07
August 07
September 07
October 07
November 07
December 07
January 08
February 08
March 08
April 08
May 08
June 08
July 08
August 08

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

PORTICO

THANK YOU.

Sheila Morrissey
sheila.morrissey@ithaka.org

http://www.portico.org

PORTICO