PORTICO

Portico Technology Overview
June 2007

Evan Owens

Chief Technology Officer

evan.owens@portico.org

www.portico.org

Portico: Business Summary

- A trusted third-party long-term preservation archive
- Initial funding by Andrew W. Mellon Foundation, JSTOR, Ithaka, and Library of Congress NDIIPP
- Mission is to preserve scholarly literature published in electronic form
- Operational in 2006 beginning with E-Journal content
  - 38 publisher participants
    - ~6100 journal titles
    - >10M articles plus new content
  - 360 library participants
- Current statistics
  - Prior to recent system upgrade, 609K articles, 3.9M files ingested
  - Currently processing 20-40K articles / day increasing to 50K / day
  - Goal is 4M to 6M articles ingested in 2007; 6M to 8M in 2008

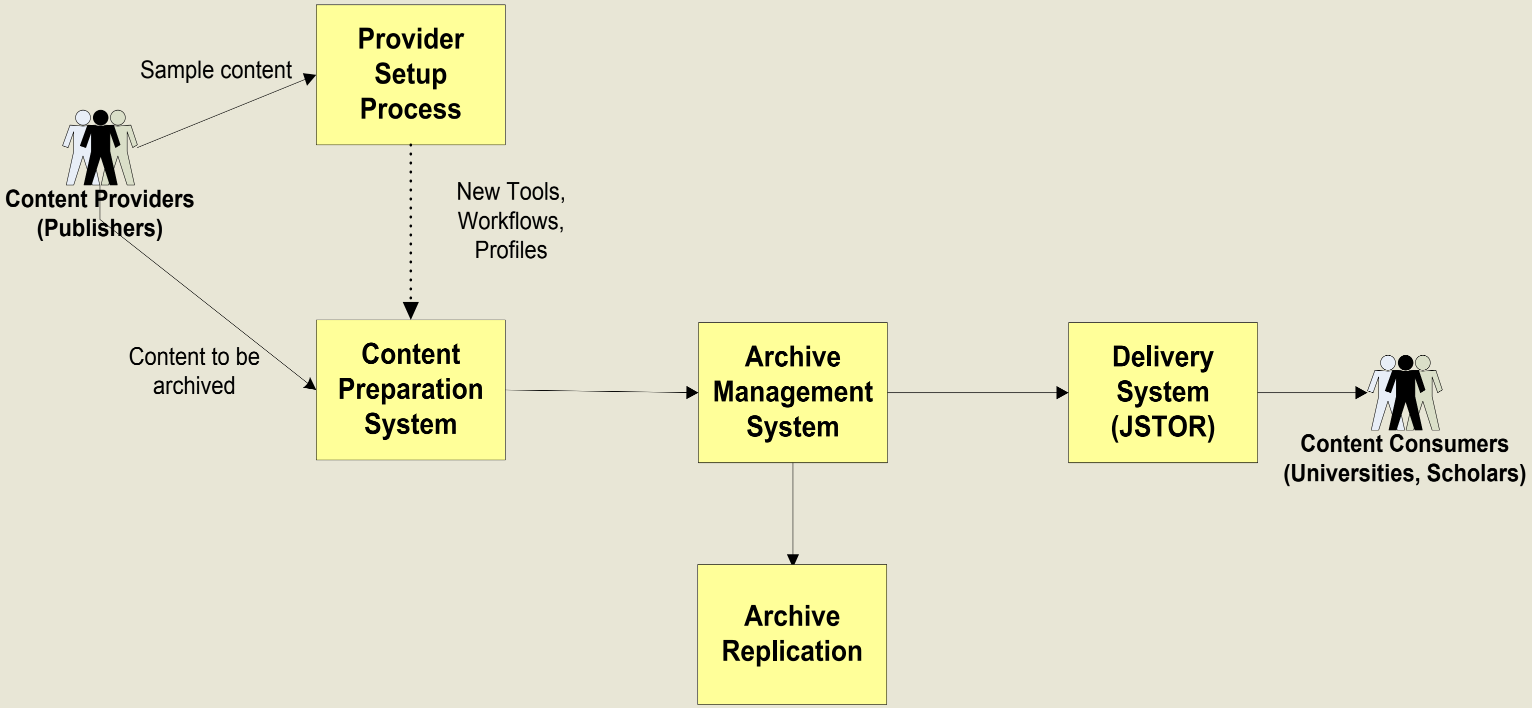
Portico: Technology Summary

- Planning began in early 2003; operational February 2006
- Key influences:
  - GDFR, PreMIS, METS, MPEG-21, ARK, OAIS
- Key technologies:
  - XML, XML schema, Schematron, JHOVE, NOID
  - Documentum, Oracle, Java, JMS, LDAP
- System design goals:
  - Pluggable tools to facilitate new providers and replacement tools
  - Configurable workflows for different content types
  - Separation of pre-processing and archive ingest
  - “Bootstrappable” archive
    - All information in the METS files
  - Metadata as first-class objects in the archive

Portico Archival Approach

- Managed preservation
- Format-based migration strategy
  - Portico Format Registry
- Preservation policies:
  - Fully supported
  - Reasonable effort
  - Byte-preserve only
- Preservation policies based on
  - Format validity
  - File format action plans and archive capabilities
  - Business rules such as publisher preferences
- Archive must also preserve supporting information
  - Required files such as DTDs and entity files
  - Documentation
  - Contracts

Systems Overview



Sample content

New Tools,
Workflows,
Profiles

Content to be archived

Provider Setup Process
Content Providers (Publishers)
Content Preparation System
Archive Management System
Delivery System (JSTOR)
Content Consumers (Universities, Scholars)
Archive Replication

Hardware Infrastructure

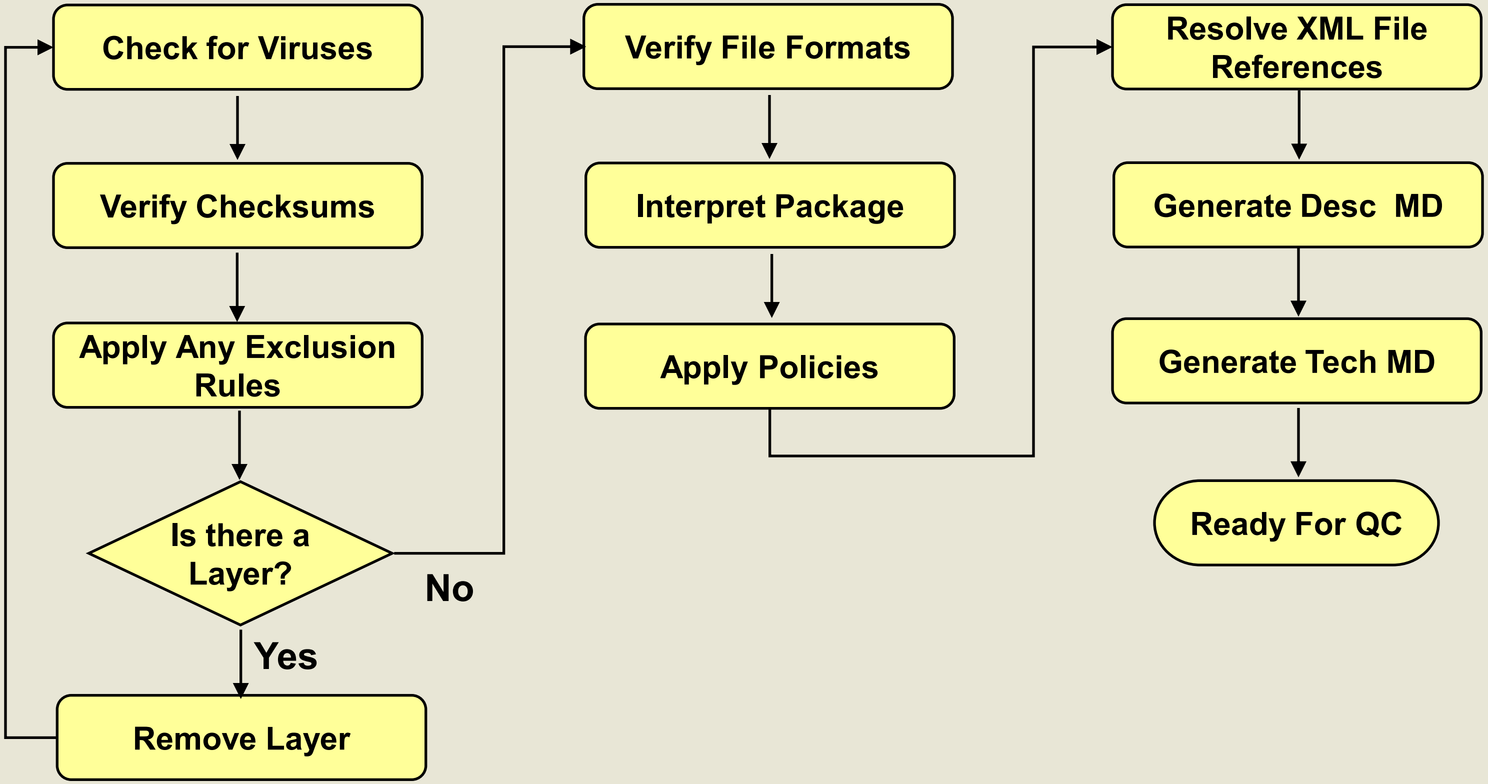
- Sun X4600 servers (Oracle servers)
- Sun v440 servers (Documentum servers)
- Sun T2000 8 core 32 thread servers (workflow activities)
- Hitachi storage frame with SATA and fiber channel disks
- OnSTOR NAS head
  - Snapshot capabilities for rapid nightly backup
- Sun L500 tape juke box
  - Writes backups and tape sets for external replication

Electronic Journal Data Issues

- Inputs
  - Per article: one text or metadata file, zero or more other files
  - Arbitrary (publisher-specific) collections of data
    - Proprietary file & directory naming conventions
    - Proprietary formats
  - Undocumented business rules hidden in the data
- Outputs
  - Content packaged in Portico METS
  - Metadata: technical, descriptive, events
  - Content restructured to Portico content model
    - Article component structure documented
  - Content normalized as per preservation plans
    - Proprietary publisher DTDs converted to NLM Archival DTD

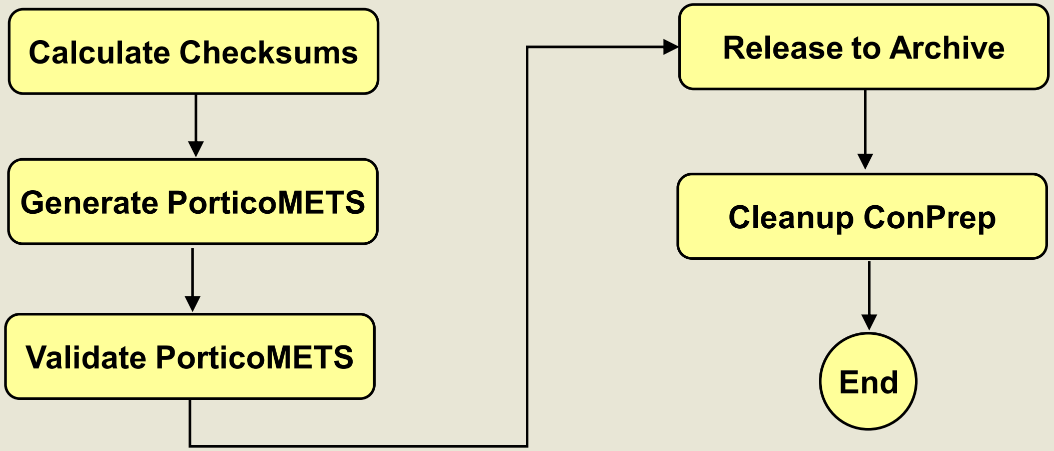
Automated Processing Workflow for E-Journal Content

(high-level summary)

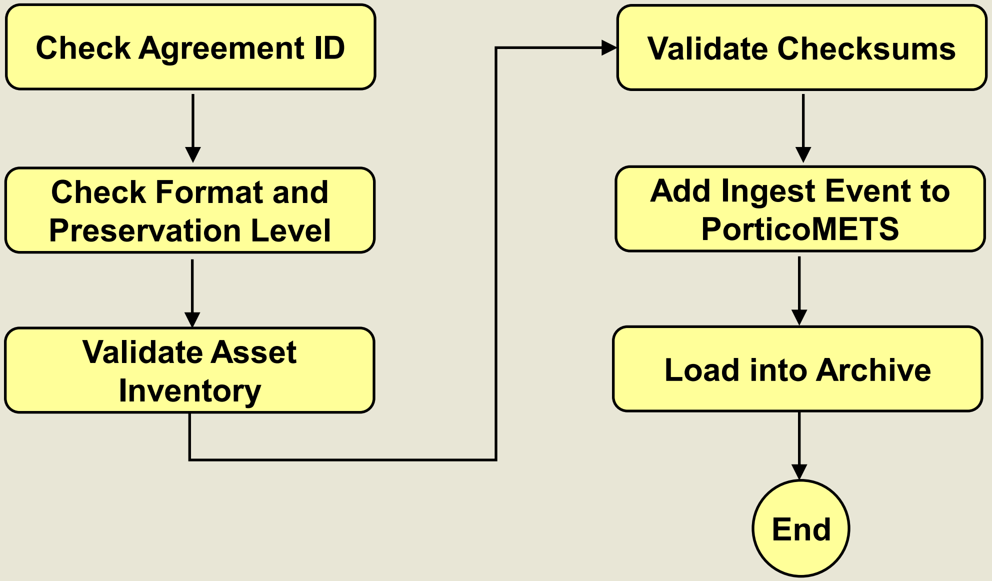


Automated Processing after QC

(for all content types)



Archive Ingest Processing



Check Agreement ID

Check Format and Preservation Level

Validate Asset Inventory

Validate Checksums

Add Ingest Event to PorticoMETS

Load into Archive

End

Key Metadata Elements

- Archival Accession Number
  - Assigned to files (physical objects)
  - Assigned to articles and components (abstract objects)
- Publisher Contract Identifiers
  - All content must map back to a publisher agreement
- Descriptive Metadata for E-Journals
  - Stored outside archived article text files
  - Policy: don't alter data in articles, but correct in metadata as needed
  - Elements that are controlled against master list:
    - ISSN
    - Publisher name
    - Title of journal
  - Elements that can be curated manually or automatically:
    - Bibliographic citation (volume, page, sequence)
    - Copyright statement
    - Article Title and Authors

Preservation Policies

- Preservation Levels:
  - Fully supported
    - Format identified and instance is valid
  - Reasonable effort
    - Format identified but not fully valid (e.g., some PDF)
  - Byte-preserve
    - Not valid or not well-formed (e.g., damaged JPEG)
    - No tool available yet for validation (e.g., MPEG)
    - Policy is not to support this format for future migration (e.g., Win32 exe
- Format-based Preservation Policies:
  - Publisher proprietary SGML / XML DTDs
    - Normalize to NLM Archival DTD
    - Pre-emptive migration
  - TIFF page images for early journal articles
    - Repackage as PDF

Portico METS Usage

- Hybrid of PreMIS, METS, MPEG-21 concepts
- Designed in parallel with PreMIS effort
- To be updated after current PreMIS review
  - To conform to currently accepted PreMIS / METS integration
- Includes fixity information
  - Currently SHA-512 checksums
- METS for METS
  - Metadata treated as an asset

Content Preparation System Components

- Workflow
  - Per content type (E-Journals, Business artifacts, Technical artifacts)
  - New and updated content
- Profiles (per provider)
  - Provider-specific rules and policies
  - Packaging rules
  - File name extract rules
- Format registry
  - List of formats known to the archive
  - Links to policy documents, technical documentation, and “required files”
- Preservation policy registry
  - What promises can the archive make for a given format?
- Tools registry & Tools service
  - What tools for which formats?
  - Where are they located?
  - How are they invoked?

Archive System Components

- Bootstrappable Archive
  - All information in METS
  - METS files and content files are sufficient to recover the archive
- Primary archive instantiation
  - Documentum CMS to manage storage of archived assets and METS
  - Oracle database with information extracted from METS
    - Cached for querying and reporting
- Additional replicas
  - On removable media in remote storage
  - In delivery/inspection system
- Integrity checks
  - 1st level: fixity information in METS files
  - 2nd level: holdings lists external to archive
    - XML schema defined manifest for content bundles

Next up on Technology Agenda

- JHOVE2
  - NDIIPP project with Harvard Library, Stanford Digital Repository
- OAI-PMH for publisher data pull
- Multiple Layers of Integrity Checks
  - Infrastructure: hardware options for self-healing, fixity, replication
  - Auditing Control Environment (ACE)
    - University of Maryland, College Park
  - Externalized metadata holdings lists to verify completeness of archive
- Tiered infrastructures
  - Archive System & Replicas
  - Low-volume and high-volume delivery
- Move to Atypon Literatum engine
  - With Aluka, JSTOR