Outline

- What is Digital Preservation?
- OAIS Reference Model
- Digital Preservation Program at CUL
- Digital Preservation Workflow
- Software and Hardware
- Archivematica Digital Preservation System
- CUL Digital Repository
- Trusted Digital Repositories
What is Digital Preservation?

- Digital Preservation is the series of managed activities necessary to ensure that digital objects can be located, rendered, used and understood in the future.
- It refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organisational change.
- The goal of digital preservation is the accurate rendering of authenticated content over time.
Why Digital Preservation Matters

- More knowledge is stored electronically
- Digital information is increasingly important to our culture, knowledge and economy
- Digital content depends on hardware and software to make it available and requires active management to ensure its ongoing accessibility
- As new technologies appear, older ones become obsolete, making it difficult to access older content
OAIS Reference Model

Open Archival Information System (OAIS) reference model provides a framework for the understanding of archival concepts needed for long term digital information preservation and access.

Submission Information Package (SIP) - information sent from the producer to the archive
Archival Information Package (AIP) - information stored by the archive
Dissemination Information Package (DIP) - information sent to a user when requested
CUL’s Digital Preservation Program

- Emerged in 2012 in response to the recognition that digitized and born-digital scholarly and cultural content needed to be carefully curated and made accessible

- Objectives:
  - Preservation processing for digitized, born-digital and delivered-digital collections following recognized national and international standards and best practices
  - Engaging in long-term preservation storage planning and management
  - Participating locally, nationally, and internationally in development of a sustainable Trustworthy Digital Repository framework
  - Guiding conversations within the Library about overall CUL policies and objectives for long-term digital preservation
CUL Long-term Preservation Responsibility

CUL is committed to preserving for the long-term:

- All master files (the highest-quality available version of the resource in an open, uncompressed file format) created as part of CUL collection digitization projects
- All master files created during external collection digitization projects that CUL assumes responsibility for preserving
- All digital content deposited into Academic Commons, our institutional repository, regardless of the file format
- Born- and delivered-digital resources that are accessioned as part of archival collections, regardless of the file format
Scope of Digital Materials

- Images
- Time-based (audio and video) materials
- Office documents in various formats
- Databases
- Email correspondence
- Websites
- Datasets
Content and Media Challenges

- Selection and sorting by content donors proves unreliable
- Privacy and confidentiality concerns
- Growing complexity of rights and access needs
- Currently more than 300 formats, multiple languages and non-roman character sets
- Long filenames/file paths (> 260 characters)
- Compressed and password-protected files
- Variety of transfer media -- hard and flash drives, CDs, DVDs, floppy disks (3.5” and 5 ¼”), ZIP disks, DV tapes, analog audio and video tapes
CUL Long-Term Archive Preservation Levels

No Preservation (Level 0)
CUL may be unable or unwilling to accept preservation responsibility. No preservation efforts except for saving a backup copy.

Bit-level Preservation (Level 1)
Basic level of preservation activity that ensures that deposited digital data (the bit-stream) is maintained in an original and valid state. Digital data will be subject to virus and fixity checks, format identification, and geographic replication. Software to view the original files is not guaranteed.

Intermediate Preservation (Level 2)
In addition to Level 1 preservation includes preservation pre-processing -- e.g., object representation information, documentation of preservation activities, creation of METS XML descriptive, technical and preservation metadata package.

Full Preservation (Level 3)
Ensures that the content of digital assets remains accessible and readable over time. In addition to Level 2 preservation, digital data will be subject to migration to archival file formats.
# NDSA Levels of Digital Preservation v2.0 - CUL

## Levels of Digital Preservation

<table>
<thead>
<tr>
<th>Functional Area</th>
<th>Level 1 (Know your content)</th>
<th>Level 2 (Protect your content)</th>
<th>Level 3 (Monitor your content)</th>
<th>Level 4 (Sustain your content)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Storage</strong></td>
<td>Have two complete copies in separate locations</td>
<td>Have three complete copies with at least one copy in a separate geographic location</td>
<td>Have at least one copy in a geographic location with a different disaster threat than the other copies</td>
<td>Have at least three copies in geographic locations, each with a different disaster threat</td>
</tr>
<tr>
<td></td>
<td>Document all storage media where content is stored</td>
<td>Document storage and storage media indicating the resources and dependencies they require to function</td>
<td>Have at least one copy on a different storage media type</td>
<td>Maximize storage diversification to avoid single points of failure</td>
</tr>
<tr>
<td></td>
<td>Put content into stable storage</td>
<td>Track the obsolescence of storage and media</td>
<td></td>
<td>Have a plan and execute actions to address obsolescence of storage hardware, software, and media</td>
</tr>
<tr>
<td><strong>Integrity</strong></td>
<td>Verify integrity information if it has been provided with the content</td>
<td>Verify integrity information when moving or copying content</td>
<td>Verify integrity information of content at fixed intervals</td>
<td>Verify integrity information in response to specific events or activities</td>
</tr>
<tr>
<td></td>
<td>Generate integrity information if not provided with the content</td>
<td>Use write-blockers when working with original media</td>
<td>Document integrity information verification processes and outcomes</td>
<td>Replace or repair corrupted content as necessary</td>
</tr>
<tr>
<td></td>
<td>Virus check all content, isolate content for quarantine as needed</td>
<td>Back up integrity information and store copy in a separate location from the content</td>
<td>Perform audit of integrity information on demand</td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>Determine the human and software agents that should be authorized to read, write, move, and delete content</td>
<td>Document the human and software agents authorized to read, write, move, and delete content and apply these</td>
<td>Maintain logs and identify the human and software agents that performed actions on content</td>
<td>Perform periodic review of actions/access logs</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Create inventory of content, also documenting current storage locations</td>
<td>Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)</td>
<td>Determine what metadata standards to apply</td>
<td>Record preservation actions associated with content and when those actions occur</td>
</tr>
<tr>
<td></td>
<td>Backup inventory and store at least one copy separately from content</td>
<td></td>
<td>Find and fill gaps in your metadata to meet those standards</td>
<td>Implement metadata standards chosen</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>Document file formats and other essential content characteristics including how and when these were identified</td>
<td>Verify file formats and other essential content characteristics</td>
<td>Monitor for obsolescence, and changes in technologies on which content is dependent</td>
<td>Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed</td>
</tr>
</tbody>
</table>
Digital Preservation Workflow

- Preservation of bit-by-bit copy of the original transfer and related documentation (media photograph, virus check report, file inventories)
- Content appraisal, selection, and arrangement (may happen any time during digital curation cycle)
- Processing of selected content with Digital Preservation software (Archivematica, BagIt)
- Transfer AIPs to Preservation Storage System
Processing/Forensic Workstations

- **FRED (Forensic Recovery of Evidence Device):**
  - Create bit-by-bit copy of the original transfer and metadata using write-blocking device and external 3.5” and ZIP drives (PC-formatted storage media)
  - Perform content analysis and selection using Forensic Toolkit
  - Upload content to staging area for preservation processing

- **Apple Mac:**
  - Create bit-by-bit copy of the original transfer and metadata (Mac-formatted storage media)
  - Upload content to staging area for preservation processing

- **IBM AT (1986) and Gateway2000 computers:**
  - Rescue data off 5 ¼” floppy disks
Software Tools

- ClamWin (ClamXav): initial virus check
- makeInventory program (based on Hashdeep): verifying content integrity (records filenames/paths, file sizes, checksums)
- Forensic Toolkit (FTK): creating disk images and content appraisal
- Aid4Mail e-Discovery Archivist: converting email to MBOX
- Database Preservation Toolkit: converting databases to XML
- Archivematica: digital preservation
Archivematica: Overview

- Open-source OAIS-compliant digital preservation system
- Compiles SIPs and produces AIPs/DIPs
- Preserves files in original formats and normalizes them to preservation/access formats
- Generates METS files containing technical, structural, descriptive, rights, and PREMIS preservation metadata
Archivematica: Content Preparation

- Content pre-processing:
  - Outsource conversion of content of analog AV media, commercially produced video DVDs, audio CDs, and mini DV-tapes to preservation formats
  - Extract data from ZIP, TAR and RAR archives
  - Convert email to MBOX
  - Convert databases to XML format

- Compiling SIPS:
  - One or multiple SIPS for each collection
  - SIP size < 1.4 TB
  - Number of files in AIP < 20,000
Metadata

- **Descriptive metadata:** makeInventory program, Archivematica, ArchivesSpace, CLIO, metadata spreadsheets from various sources
- **Technical metadata:** makeInventory program, Forensic ToolKit, Archivematica
- **Preservation metadata:** Archivematica
- **Rights metadata:** Archivematica, Hyacinth
CUL Digital Repository

- CUL established designated Isilon local storage space for acquired digital content (/digital/).
- Files stored in this space are in various stages of transfer, appraisal, processing, and preservation.
- Files move between storage locations during digital curation cycle.
Digital Content Flow and Storage Space Allocation

- **dark-archive**: space for content that needs long-term preservation, but requires infrequent or no access during extensive periods of time.
- **working**: space for digital materials that undergo triage, appraisal or processing.
- **ingest**: space for digital materials awaiting for Intermediate Preservation Processing or preserved at Bit-level. Some processing and metadata work might still be in progress.
- **preservation**: permanent location for digital materials that have gone through appropriate digital preservation processing. The files may also be used for preparing access copies if those do not already exist.
- **access**: permanent location for access copies. Files in this directory are referenced by Fedora/Hyacinth when access is provided.
COLUMBIA UNIVERSITY LIBRARIES DIGITAL PRESERVATION ARCHITECTURE

Hyacinth

- Generation/remediation of:
  - Descriptive metadata
  - Technical metadata
  - Rights metadata
  - Structural metadata

Columbia University Libraries Digital Preservation Architecture

Digitization Projects

Inst. Repository / Data Sets

Born-Digital Archives

Preservation Repository Services

Auditing
Access management
Integrity checking
Content stabilization/ transformation
Format migration

Fedora Repository

EMC Isilon Cluster (at Columbia)
900 TB

EMC Isilon Cluster (Syracuse, NY)
900 TB

Cloud

Publishing Targets:
- Digital Library Collections Portal
- Special Custom Sites
- etc.
Digital Storage

- Total replicated Isilon storage capacity: ~900 TB

- New content added in 2019 (through 10/2019): ~ 93 TB
  (includes e.g., AMI/TBM, Bob Fass audio archive, Muslim World Manuscripts)

- Remaining Isilon storage (as of 11/2019): ~139 TB

- Transition from Isilon hardware-based storage to a full cloud storage platform: 2020-2021 (with LIT)

- Major improvements in data protection, replication, cost and availability.
## Trusted Digital Repositories

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>CORE</th>
<th>EXTENDED</th>
<th>FORMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization(s)</td>
<td>CoreTrustSeal Board</td>
<td>DIN (German Institute for Standardization)</td>
<td>ISO (International Organization for Standardization)</td>
</tr>
<tr>
<td>No. of Requirements</td>
<td>16</td>
<td>34</td>
<td>100+</td>
</tr>
<tr>
<td>Standards</td>
<td>CoreTrustSeal</td>
<td>DIN 31664</td>
<td>ISO 14721 (OAIS), ISO 16363</td>
</tr>
<tr>
<td>Audit Process</td>
<td>Self-assessment + independent peer review (2)</td>
<td>Self-assessment + independent peer review (2)</td>
<td>ISO certified audit with accredited auditors</td>
</tr>
<tr>
<td>Certification lifespan</td>
<td>3 years</td>
<td>Indefinite</td>
<td>3 years</td>
</tr>
<tr>
<td>No. of Certified Repositories</td>
<td>149</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

Major assessment areas are the same for all three levels and include:

- Organizational Infrastructure
- Digital Object Management
- Technologies, Technical Infrastructure, and Security
Questions?