Automated Digital Processing at the Bentley Historical Library

Nancy Deromedi and Michael Shallcross

Bentley Historical Library • University of Michigan • 1150 Beal Avenue • Ann Arbor, MI 48109-2113 U.S.A.

Institutional Context

- Bentley Historical Library established in 1935 by the University of Michigan Regents to:
  - Serve as official archives of the University of Michigan
  - Document the history of the state of Michigan and the activities of its people, organizations, and voluntary associations
- Comprised of four divisions:
  - University Archives and Records Program
  - Michigan Historical Collections
  - Reference and Access Services
  - Digital Curation (established 2011)
- No in-house IT department

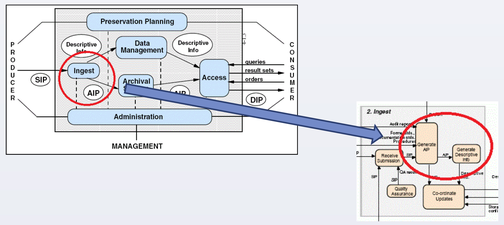
Managing Digital Content at the Bentley

- First major deposit of digital material in 1997
  - Former University President James J. Duderstadt's digital desktop
- MeMail Project (2010-2011):
  - Funded by the Andrew W. Mellon Foundation
  - Provided the library with resources to establish a workflow, policies, and resources for the ingest and processing of archival email
  - See: http://e-records.chrisprom.com/memail-project-guest-post-by-mike-shallcross/
- After MeMail, a similar solution was needed for mixed digital content (Office documents, PDFs, audio, video, images, etc.)

A Definition of "Digital Processing"

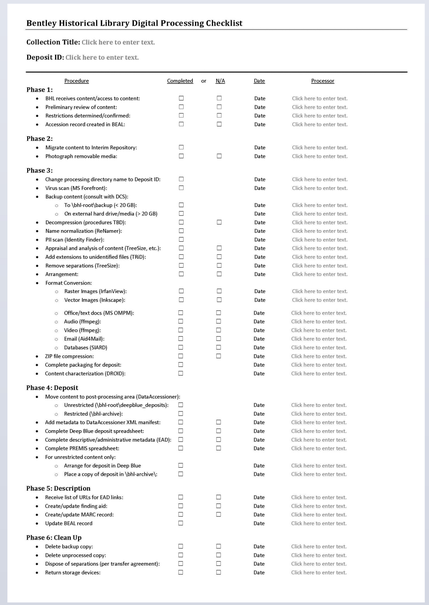
- Traditional archival functions:
  - Appraisal
  - Arrangement (only if necessary)
  - Description
- Additional steps to ensure the authenticity, integrity, and security of content
  - Assume intellectual control of content
  - Perform preservation actions
  - Establish integrity of materials
- Goal: generation of Archival Information Package (AIP)

Digital Processing in terms of OAIS



Developing a Digital Processing Workflow

- Manual workflow established in 2011:
  - 40 discrete steps from ingest through deposit and description
  - Required:
    - Operation of numerous software applications
    - Following strict conventions for naming and saving various log files
    - Creation of metadata records by hand
    - Tracking processing status with a form



Bentley Historical Library Digital Processing Checklist

Collection Title: Click here to enter text.

Deposit ID: Click here to enter text.

Procedure	Completed	or	N/A	Date	Processor
Phase 1:					
• BHL receives content/access to content:	☐		☐	Date	Click here to enter text.
• Preliminary review of content:	☐		☐	Date	Click here to enter text.
• Restrictions determined/confirmed:	☐		☐	Date	Click here to enter text.
• Accession record created in BEAL:	☐		☐	Date	Click here to enter text.
Phase 2:					
• Migrate content to Interim Repository:	☐			Date	Click here to enter text.
• Photograph removable media:	☐		☐	Date	Click here to enter text.
Phase 3:					
• Change processing directory name to Deposit ID:	☐			Date	Click here to enter text.
• Virus scan (MS Forefront):	☐			Date	Click here to enter text.
• Backup content (consult with DCS):					
○ To \bhl-root\backup (< 20 GB):	☐			Date	Click here to enter text.
○ On external hard drive/media (> 20 GB)	☐			Date	Click here to enter text.
• Decompression (procedures TBD):	☐		☐	Date	Click here to enter text.
• Name normalization (ReNamer):	☐			Date	Click here to enter text.
• PII scan (Identity Finder):	☐			Date	Click here to enter text.
• Appraisal and analysis of content (TreeSize, etc.):	☐		☐	Date	Click here to enter text.
• Add extensions to unidentified files (TRID):	☐		☐	Date	Click here to enter text.
• Remove separations (TreeSize):	☐		☐	Date	Click here to enter text.
• Arrangement:	☐		☐	Date	Click here to enter text.
• Format Conversion:					
○ Raster Images (IrfanView):	☐		☐	Date	Click here to enter text.
○ Vector Images (Inkscape):	☐		☐	Date	Click here to enter text.
○ Office/text docs (MS OMPM):	☐		☐	Date	Click here to enter text.
○ Audio (ffmpeg):	☐		☐	Date	Click here to enter text.
○ Video (ffmpeg):	☐		☐	Date	Click here to enter text.
○ Email (Aid4Mail):	☐		☐	Date	Click here to enter text.
○ Databases (SIARD)	☐		☐	Date	Click here to enter text.
• ZIP file compression:	☐			Date	Click here to enter text.
• Complete packaging for deposit:	☐			Date	Click here to enter text.
• Content characterization (DROID):	☐			Date	Click here to enter text.
Phase 4: Deposit					
• Move content to post-processing area (DataAccessioner):					
○ Unrestricted (\bhl-root\deepblue_deposits):	☐			Date	Click here to enter text.
○ Restricted (\bhl-archive):	☐			Date	Click here to enter text.
• Add metadata to DataAccessioner XML manifest:	☐			Date	Click here to enter text.
• Complete Deep Blue deposit spreadsheet:	☐		☐	Date	Click here to enter text.
• Complete description/administrative metadata (EAD):	☐		☐	Date	Click here to enter text.
• Complete PREMIS spreadsheet:	☐		☐	Date	Click here to enter text.
• For unrestricted content only:					
○ Arrange for deposit in Deep Blue	☐			Date	Click here to enter text.
○ Place a copy of deposit in \bhl-archive\;	☐			Date	Click here to enter text.
Phase 5: Description					
• Receive list of URLs for EAD links:	☐		☐	Date	Click here to enter text.
• Create/update finding aid:	☐		☐	Date	Click here to enter text.
• Create/update MARC record:	☐		☐	Date	Click here to enter text.
• Update BEAL record	☐			Date	Click here to enter text.
Phase 6: Clean Up					
• Delete backup copy:	☐		☐	Date	Click here to enter text.
• Delete unprocessed copy:	☐		☐	Date	Click here to enter text.
• Dispose of separations (per transfer agreement):	☐		☐	Date	Click here to enter text.
• Return storage devices:	☐		☐	Date	Click here to enter text.

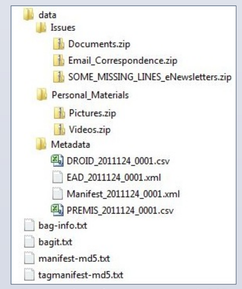
Original manual processing checklist

- Problems with manual approach:
  - Highly labor intensive
  - Numerous opportunities for inconsistencies and user error
  - Difficult to train staff
  - Requires moderate to advanced technical skills

Automating the Processing Workflow

- Developed the AutomatedProcessor (AutoPro) to permit archivists to focus on traditional functions:
  - Successful proof of concept in early 2012; deployed in production environment in June 2012
  - Guides archivists through a standardized workflow to ingest, process, and describe digital content
  - Inspired by the Archivematica Digital Preservation System and the National Archives of Australia's Digital Preservation Software Platform (DPSP)
  - Successfully processed over 60 deposits (with volumes of up to 45,000 files/100 GB) to date
- Features of AutoPro:
  - Comprised of 33 Windows CMD.EXE shell scripts that control more than 20 applications and various command line utilities
  - Moves content through a 12 step workflow; requires user approval to advance to each step
  - Archivists may stop at any point in the workflow and resume work at a later time
  - Complete audit trail; documents all modifications and preservation events during AIP generation:
    - Log files for all operations
    - PREMIS preservation metadata file
  - Produces Dublin Core metadata for batch upload to DSpace repository

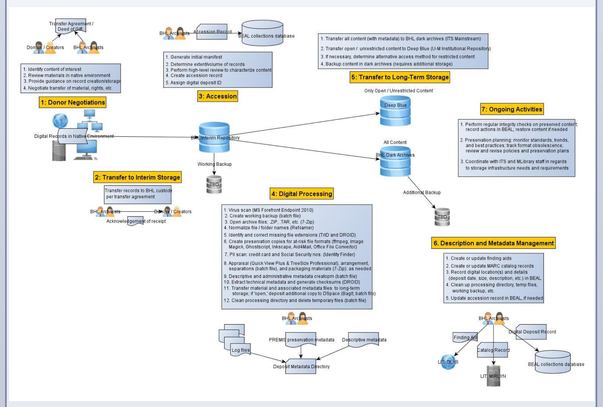
Windows Command Prompt Interface



Example of a processed AIP in Dark Archive

Unrestricted content in DSpace

End-to-End Digital Preservation Workflow



Processing Workflow: Steps and Software

Workflow Step	Definition	Software/Utility
1. Virus scan	Scan for viruses and malware	MS Forefront Endpoint Protection
2. Create temporary backup	Establish restore point in case of data loss	Windows utility
3. Open archive files (.ZIP, .TAR, etc.)	Extract content to directories	7-Zip
4. File and folder name normalization	Replaces spaces and non-alphanumeric ch. with underscores	ReNamer
5. Identify missing file extensions	Identify and add file extensions as needed	TrID and DROID
6. Create preservation copies	Use set conversion strategies to create a preservation copy of original bitstream	ffmpeg, Ghostscript, Inkscape, Aid4Mail, Image Magick, Office File Convertor
7. PII (credit card and Social Security number) scan	Identify sensitive info.; redact where possible or apply access restrictions	Identity Finder DLP Endpoint
8. Appraisal and arrangement	Review files, separate unneeded content, arrange and package (in uncompressed .ZIP files) if necessary	Quick View Plus (file viewer), TreeSize Professional (file manager), 7-Zip, and batch files
9. Descriptive and administrative metadata creation	Record descriptive & admin. metadata; stored in Dublin Core XML manifest and EAD file, respectively	(Manual data entry)
10. Extract technical metadata	Generate DROID report with item-level checksums	DROID
11. Transfer content (with metadata) to long-term storage	All content to dark archive; unrestricted material also deposited in DSpace	BagIt
12. Clean up	Clean processing directory; delete temp files	Windows utilities

Moving Forward...

With continued development, the Bentley Historical Library hopes to produce an open source tool for the archival community at large.