ELECTRONIC RECORDS INGEST

Examples & Models
PRESENTERS

Veronica Martzahl
Massachusetts Archives
Electronic Records Archivist

Mark Myers
Texas State Library and Archives Commission
Electronic Records Specialist
ACKNOWLEDGEMENTS

This webinar is made possible by a grant from the National Historical Publications & Records Commission (NHPRC)
TERMS AND MA EXAMPLES

Veronica Martzahl
Massachusetts Archives
n. ~ In the Open Archival Information System (OAIS) model, processes related to receiving information from an external source and preparing it for storage.¹

REPOSITORY?

• Your Archives

• Network Storage

• Surrogate Repositories

• Digital Preservation Repository
EXAMPLES OF DIGITAL PRESERVATION REPOSITORIES
PAIMIS AND PAIS

ISO 20652: Space Data and Information Transfer Systems - Producer-Archive Interface - Methodology Abstract Standard

Under development: PAIS – Producer-Archive Interface Specification

Guidance for establishing your submission agreement and getting the information you need for the Submission Ingest Package (SIP) you need for ingest.
Ingest = Getting it into our custody
Repository = Managed Network Storage

Steps

• Transfer
• Stabilization
• Monitoring
TRANSFER

Mechanism?
• External Hard Drive
• FTP (File Transfer Protocol)

Requirements?
• File types
• Metadata
STABILIZATION

• Virus scan
• Checksum(s)
• File migration
• Documentation
  • Read Me file
  • Basic Arrangement and Description
LEVEL OF DESCRIPTION

1. Are the files restricted?

2. Is individual access to a file appropriate/needed for reference or other access services?

3. Are the files going into a repository with a public interface or are they going into a dark archive?

4. Is the office or collection high profile and/or does it have a history of intensive processing?
MONITOR

- Make sure storage media is up to date
- Check checksums
- Review file formats
Ingest = Into a Preservation Repository
Repository = Fedora

Steps

- Transfer
- Stabilization
- Arrangement and Description
- Monitoring
TRANSFER

Mechanism?

• TAPER – Tufts Accessioning Program for Electronic Records

Requirements?

• Transfer form
• Xythos dropbox
TRANSFER FORM

Tufts Accessioning Program (TAPER) Home

This is the homepage of all the TAPER tools.

Transfer Agreement Form

View Preservation Rules
PRE-INGEST INFORMATION

Office of Origin
Authorized representatives
Creator
Producer
Email contact information
Record type
Dates
Copyright

Access
Formats
Arrangement and Description
Retention Period
Descriptive Standard
Respect de Fonds
Connection to dropbox
STABILIZATION

• Virus scan
• Checksum(s)
• File migration
ARRANGEMENT AND DESCRIPTION

CIDER

• Archival Collection Management System
• Exports metadata to create FOXML object for FEDORA Repository
• Individual file vs. Aggregate description (Tufts Digital Library or Dark Archive)
MONITORING

- Multiple FEDORA instances
- CIDER metadata
DEVELOPMENTS

- Hydra administrative interface to streamline the ingest process
- Admin interface also allows other library professions to do some of the metadata work for the archives (ie subject cataloging of Senior Theses)
IN SUMMARY

- **Goals**
  - Get the files into our custody
  - Document, document, document
  - Keep the files safe, regardless of the environment we are working in
    - Able to proved what you have is what you received
    - Able to show no deterioration has occurred
    - Able to find and access the files
KENTUCKY AND TEXAS EXAMPLES

Mark Myers
Texas State Library and Archives Commission
KDLA E-ARCHIVE

Ingest = Locating the records & Getting them into the repository

Repository =

• 1996-2008 = Managed Network Storage &
• 2008 = 2012 = Dspace &
• 2012-Present = Preservica

Steps =

• Acquisition (locating the records)
• Transfer
• Arrangement & Description
• Monitoring = Preservation
ACQUISITION METHOD #1: DIG IT UP YOURSELF!

Harvesting from agency websites.

- Manual process
- Have to hunt around on site
- Prioritized list of agencies
  - Outgrowth of state publications program
  - Followed retention schedule process
- Acquired tools along the way
WEB HARVESTING/COLLECTING TOOLS

Grab-A-Site (Blue Squirrel)


  Pros:
  - Allowed for crawling an entire website and grabbing files of a specific type (.doc, .jpg, .pdf)
  - Also used to archive whole websites (Governor, Lt. Gov., First Lady)

  Issues:
  - Not automated
  - No metadata – websites captured as HTML and other formats
  - Not sure that it’s still being maintained. Version 5.0 has been out for over 10 years.

Archive-It ([www.archive-it.org](http://www.archive-it.org))

- KY - Became full partner in 2010; TX - partner since 2007

  Pros:
  - Allows for AUTOMATED harvesting and metadata addition/collection
  - Can navigate the morass that is the Internet: Content Mgmt systems, Social Media, etc

  Issues:
  - Can’t harvest for just objects (pdf, doc, jpg) – what if you don’t want the entire website?
  - Records are located on Archive-It site

Preservica ([http://preservica.com/](http://preservica.com/))

- Allows for web harvesting directly into repository
  - New feature in current version of Cloud Edition
  - Also allows for capture from known FTP sites (Enterprise Edition)

- Uses Heritrix web crawler

  Issues:
  - Same as Archive-It – records are wrapped up in WARC file
WEB HARVESTING/COLLECTING TOOLS

Download Them All

- Browser extension for

- Allows for grabbing all the files of a certain type off a single page

- Filter by file type
  - Documents (pdf, odf, doc)
  - Customizable – can add to filters

- Allows for renaming

- Issues:
  - Single page
  - No metadata, just object
WEB HARVESTING/COLLECTING TOOLS

YTD Video Downloader

http://download.cnet.com/YTD-Video-Downloader/3000-2071_4-10647340.html

- Allows for identification and capture of media files on a page
- Download and convert audio/video to other formats
- Allows for starting/stoping
- Freeware & Paid (annual subscription)
  - Auto-conversion is a PRO feature
  - Free version will convert from already downloaded files

Issues:
- Single page, not whole site
- No metadata, just object
- Legality???
ACQUISITION METHOD 2: Talk to Agencies (or “Records Producers” in OAIS-speak)

Communicate with agencies:

- Ask for electronic records!
- Coordinate harvesting
  - Tell us where records are
  - Remove impediments to harvesting
  - Tell us when records will be taken down (preferably before)
- Coordinate transfers
  - Send us the records
  - Policies/Procedures
  - Method
  - Metadata
ACQUISITION METHOD 3: Make (get) them do it for us

• Direct deposit into DSpace
  • Good for single/small transfers
  • Homogenous documents
    • Serials
    • Minutes
    • Pubs
  ▪ Get’s the creator to add metadata through templates
  ▪ Archivist can approve/reject/edit
  ▪ Issues:
    • Never quite got off ground
    • Not good for large transfers
TRANSFER 1 - INSTRUCTIONS

Instructions = Policies & Procedures

• New Administrative Regulation for State Publications
  ▪ Allow for electronic transfer (instead of paper)
  ▪ Mandates PDF format (state standard for published material)

• Records Transfer Guidance
  ▪ Coupled with paper process
  ▪ Outlined the methods
  ▪ Outlined “desired” formats

## Recommended Data Formats for Preservation Purposes in the KDLA Digital Archive

http://kdla.ky.gov/records/recmgmtguidance/Documents/File%20Format%20table.PDF

<table>
<thead>
<tr>
<th>Media</th>
<th>High Confidence</th>
<th>Medium Confidence Level</th>
<th>Low Confidence Level</th>
<th>Notes/Comment</th>
</tr>
</thead>
</table>
| Text           | - Plain text (encoding: US ASCII, UTF-8, UTF-16 with BOM)  
                - PDF/A-1 (*.pdf)  
                - XML (XSD/XSL/XHTML, etc.; with included or accessible schema and character encoding explicitly specified) | - Plain text (ISO8859-1 encoding)  
                - PDF (*.pdf) (embedded fonts)  
                - Rich Text Format (*.rtf) version 1.x  
                - OpenOffice (*.sxw)  
                - Microsoft Word (*.doc)  
                - WordPerfect (*.wpd)#  
                - HTML 4.x (include a DOCTYPE declaration)  
                - SGML | - PDF (external font)  
                - All other text formats not listed here  
                - DIVX (alternative format to PDF. Uses a different compression to make a smaller file.  
                  Published standard. Created by AT&T; owned by producers of MsSID GIS format.  
                  Used by USGS and other GIS and Washington State | |
| Raster Image   | - TIFF (uncompressed)  
                - PNG (*.png)  
                - JPEG (raw)? | - BMP (*.bmp)  
                - JPEG/JFIF (*.jpg)  
                - JPEG2000 (prefer uncompressed) (*.jp2, *.jpx)  
                - TIFF (CCITT Group 3/4, JPEG, PackBits compression) | - MsSID (*.sid)  
                - TIFF (with LZW compression or in Planar format)  
                - GIF (*.gif)  
                - FlashPix  
                - Photoshop (*.psd)  
                - All other raster image formats not listed here | - “Raw” JPEG are those images that have not been resized.  
- Depends on compressio n format  
- Uncompressed |
| Audio          | - AIFF (uncompressed) (*.aiff)  
                - WAVE (LPCM only) (*.wav) | - Standard MIDI (*.mid, *.midii)  
                - Windows Media Audio (*.wma)  
                - MP3 (MPEG 1/2, Layer 3) (8.mp3)  
                - SUN Audio (uncompressed) | - AIFFC (*.aiffc)  
- NeXT SND (*.snd)  
- RealNetworks 'Real Audio' (8.ra, *.rm, *.ram)  
- WAVE (compressed) (*.wav)  
- MP3 is a non-documented compressed version of MPEG  
- the bare MPEG is open (v. 1 & 2 are |
TRANSFER 2 – METHODS

• Email
  • Good for small number of files (attachment)
  • Limitation by (total) file size
• On Media (CD/DVD, Flash Drives, External Hard Drive)
• FTP (File Transfer Protocol) or Website
  • (FTP) Good for incredibly large files
  • (Website) Good for large number of files already out there
• Direct Deposit
  • Across the LAN (GIS Snapshots)
  • DSpace
Preservica SIP Creator

- Used to ingest files into Preservica
- Allows for pulling together a SIP package
Preservica SIP Creator

- Apply Integrity Checks
- Attach metadata to the package
ARRANGE & DESCRIBE: Place files on file server (local storage)
Accession metadata!!!!
KDLA’S ELECTRONIC RECORDS
ACCESSION FORM

Metadata captured:

- **Agency Info** - Includes Record Group number which links out to other (paper) accession databases
- **Accession number** – “E” (for electronic)/year/000-999
- **Method of Transfer** – Optical disk, Email, Download from website, FTP. Download site url if applicable
- **Total number of files**
- **Volume in MB**
- **For each series:**
  - Series Number
  - Disk (1 of x)
  - Folder name
  - Number of Files
  - Format (PDF, DOC, JPG, TIF)
  - Date span
- **Technical Contact info** – person responsible for electronic record
- **Records Officer**
- **Other technical/preservation metadata based on file type**
  - Only PDF and Digital images
KDLA’S ELECTRONIC RECORDS ACCESSION FORM

Issues:

- All manual – not automated
  - Only internal – we have to fill out
  - Manual form on website, but staff still have to enter into database
  - No extraction of metadata from files/objects

- Doesn’t capture file names – only total number of files
  - No fixity information (checksums)
  - Only sight verification – open folders and count number of files

- Mixed media an issue – folders with multiple file types

- Many times don’t have technical contact info

- Technical/preservation information is often incomplete
  - Again, manual – no validation/extraction of technical information

- Mirrors Paper system
ARRANGEMENT & DESCRIPTION: Enter DSpace
Cropland Data Layer Maps for the Commonwealth of Kentucky [2008]

Agriculture, Kentucky Department of; U.S. National Agricultural Statistical Service

URL: http://hdl.handle.net/10602/11466
http://daasnp1.state.ky.us:8080/xmlui/handle/10602/11466
Date: 2008

Abstract:
Downloaded from the U.S. National Agricultural Statistical Services in KML and PDF format

Show full item record

Dspace Full Item Record
MONITOR = PRESERVATION

Preservica preservation Processes

- Virus Scan
- Characterization and Format validation
- Normalization & Creation of DIP’s
  - Original and Derivative managed together
- Integrity Checks & Monitoring
- Migration actions
SUMMARY

• Main Points:
  • Find the records
  • Bring them under control = Instruction/Guidance
  • Apply appropriate metadata
    • Accession info
    • Administrative
    • Preservation/technical
  • Make the records accessible (arrangement, description)
  • Maintenance/Preservation
SUMMARY

• Finding the records – communication is the key
  • Know where the records are
  • How to get the records to archive

• Automate the process as much as possible
  • Metadata capture
  • Harvesting
  • Making systems talk to each other
SUMMARY

• Keep it Simple
  • Agencies aren’t going to repackage, rename, rearrange, use special tools/process

• Keep it secure
  • But – need to ensure integrity through the transfer process
  • Make sure you received what was sent
  • Objects haven’t changed
CONTACT INFORMATION

Veronica Martzahl
veronica.martzahl@sec.state.ma.us

Mark Myers
mmyers@tsl.texas.gov
QUESTIONS & COMMENTS

It's QUESTION TIME!!
WRAP-UP

- Post-webinar evaluation will automatically open in your web browser when you exit the session.

- Next SERI Educational Webinar is November 5 @ 2:00pm
  - Topic: Electronic Records Storage

- Complete webinar schedule is available on CoSA’s website:
  http://www.statearchivists.org/CoSA_Webinars.htm

- All webinar slides available from the SERI webinar page:
  http://www.statearchivists.org/seri/STEP/SERI_Educational_Webinars.htm