UAB Libraries
Digital Preservation Workflows

Grayson Murphy – gmurphy@uab.edu
Digital Preservation Librarian
About University of Alabama at Birmingham (UAB)

- Established in 1969
- Public research university with 22,000+ students
- Located in Birmingham, Alabama, United States
- One of the largest Academic Medical Centers in the United States.
UAB Libraries – Digitization and Digital Preservation Department

- Formed in 2021 (previously the digital unit within cataloging)
- First team to focus on digital preservation at UAB Libraries
- Four full-time team members + Student Workers
- Digitize and preserve various content:
  - Historical collections
    - Print and A/V materials
  - University publications
    - Print and born digital
  - Faculty and student work
    - Theses and Dissertations
    - Publications

Head of DnD

Digital Imaging Manager

Digital Preservation Librarian

Digital Curation Librarian
Digitized Print Materials Workflow
Table of contents

1. Questions that Guided Workflow Creation
2. Imaging
3. Generating SIPs
4. Transfer to Preservation Staging Server
5. Quality Control and Ingest into Preservation Storage
6. Sticking Points
7. Lessons Learned
Workflow Questions – Adapted from slides created by Luke Menzies

• How do we implement OAIS reference model?
  • Ingest most crucial step
  • AIP as the building block
  • Ensure process is well-documented, consistent, and standards-compliant
Workflow Questions — Adapted from slides created by Luke Menzies

- How do we structure our AIPs?
  - Item vs. Collection
  - manifest.csv
  - metadata.csv
  - BagIt and Tar It
Workflow Questions – Adapted from slides created by Luke Menzies

• How do we minimize risk?

<table>
<thead>
<tr>
<th>Risk</th>
<th>&quot;Rule&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data loss during processing</td>
<td>Process objects on the <strong>same machine</strong> where they were created</td>
</tr>
<tr>
<td>Data loss and date-time metadata during transfer</td>
<td>For <strong>physical</strong> transfers, Tar/ Zip and run fixity checks on both sides</td>
</tr>
<tr>
<td>Something goes wrong when updating AIPs</td>
<td>Only store <strong>minimal, immutable</strong> metadata within AIPs</td>
</tr>
<tr>
<td>Something goes wrong during processing</td>
<td>Process in <strong>batches</strong> and keep automations <strong>modular</strong></td>
</tr>
<tr>
<td>An imperfect AIP is stored which is not accessible or usable</td>
<td>Incorporate &quot;<strong>stop</strong>&quot; or &quot;<strong>slow down</strong>&quot; points in the workflow, with human eyes</td>
</tr>
</tbody>
</table>
Step 1 - Imaging

- Digital Services Lab images object
  - Flatbed scanners + DT Atom
  - Generates tiffs for preservation master copies
  - Creates derivatives for access copies (not put into SIPs)

- Keeps track of imaging in shared processing log
  - Each object has its own line in a spreadsheet to track its lifecycle through the workflow
  - Processing log sections include:
    - Accessioning
    - Imaging
    - Preservation
    - DIPStore (storage for access master copies)

DT Atom is used to image rare / fragile materials not suitable for a flatbed scanner.
Part of the Imaging section of the Processing Log
Step 2 – Generating SIPs

- DnDprepack.py
  - Generates minimal metadata csv that is loaded into the next script
  - User inputs local ID after CSV created

GUI for DnDprepack.py
Step 2 – Generating SIPs

- SIPmaker.py
  - Writes minimal metadata files from csv, generates manifest files, bags and tars object

GUI for SIPmaker.py
Step 3: Transfer to Preservation Staging Server

- Digital Services Lab transfers SIPs from lab computer to network attached server (NAS)
  - Uploads to preservation share on Synology NAS
  - Assures integrity with fixity checks run on both sides of transfer
Step 4: Quality Control and Ingest into Preservation Storage Locations

- Pre-upload Quality Control
  - Use 7zip to open .tar archive and check if:
    - SIP structure is correct
    - apttrust-info.txt file has correct S3/ Glacier geographic destination
    - Minimal metadata csv is populated correctly

- Upload items using simple CLI python script
  - Utilizes Boto3 SDK for AWS
  - User inputs:
    - Demo or Production environment
    - Personal keys
    - Path of items to upload

Preservation section of the processing log
Sticking Points

• How to ensure safety of physical items
  - In a lab only accessible by scanning campus ID
  - Dehumidifier running 24/7
  - Student workers trained in safe handling practices

• Where to put scanned images during imaging
  - Put scanned images on external hard drive connected to lab computer rather than on lab computer
  - Digital Imaging Manager can always access student work as opposed to it being locked in a student’s user account
  - Hard drives can be moved around so content is not locked to one specific computer

• Preserve derivative tiffs or not
  - Pros:
    - Preserves post-processing of images
  - Cons:
    - Effectively doubles the size of preservation packages
      - Higher cost and environmental impact long term
    - Complicates Digital Services Lab workflows
  - Solution: Do not preserve derivative tiffs in preservation packages but put them in short term DIPStore on local NAS share
Lessons Learned

• Be realistic as to what people executing the workflow can/will do
  • You can build the most ideal system but if users won’t follow through then it’s pointless
  • Keep the bar low for "future us"

• Iterate through the process and expect to find flaws

• Be open to reworking even the most established elements
  • Had to repackage 5 TB of data (down to 2.5) when we removed the derivative tiffs

• Be open to feedback from everyone involved
  • New people bring new ideas
Questions?
Preserving Legacy Digital Collections Workflow
Table of contents

1. Background
2. Scope
3. Goals
4. Questions to Address
5. Master Workflow
6. Project Plan
7. Lessons Learned
Background

• First Digital Collection created in 2006
  • 30+ more created until DnD department established
  • CONTENTdm (2006-2023)
  • Alma Digital (Present)

• Content backed up on 15 hard drives over the years
  • Drobo NAS purchased in 2017 and copies of hard drives put on Drobo

• 2023
  • Digital Preservation Librarian is hired
  • Drobo goes out of business
    • No longer updating software or supporting customers
    • Only way to access Drobo admin is through their proprietary software which is unreliable
Scope

- 3.6 TB of data
  - Oral Histories
  - Images
  - Videos
  - Digitized print materials
- 30+ digital collections
- Content represents nearly 20 years of work
Questions to Address

• Do we structure preservation packages the same way?
  • First time putting things into preservation that were not generated in the digital services lab
  • File hierarchy, file names, etc.

• How to keep track of the process?
  • Once the workflow is created, how do we add checkpoints?

DnD File Naming Convention
Goals

- Organize, package, and put all items represented in current digital collections into preservation
  - No uniform organization to collection folders on Drobo
  - Each collection has its own file organization
- Discover and note materials that are not currently in digital collections
  - Do we keep it or get rid of it?
  - Is it worthy of adding to a collection?
DPL = Digital Preservation Librarian
DCL = Digital Curation Librarian
IROL = Institutional Repository Outreach Librarian

**Appraise**
- DPL assess collection
- DPL confer with Digital Imaging Manager about what is on Drobo in relevant collection

**Plan**
- DPL inform DCL / IROL about what is in on Drobo for that collection
- DPL confer with DCL/IROL about what they want in DIPStore from collection
- Decide what data to keep

**Organize (DPL)**
- Delete files not being kept if interfering with reorganization
- Create Inventory Log
- Reorganize file structure
- Connect contents to presentation platform (if applicable)
- Assign UUIDs
- Rename Files
- Zip staging items and upload to staging

**Preservation**
- Create Processing log and transfer relevant info from Inventory Log
- Figure out Local ID format with relevant parties
- Package items and update Processing Log
- Upload to APTrust
- Transfer to Synology

**Documentation**
- DPL add Collection Object Structure and Local ID decision to Confluence
- DCL or IROL add UUIDs to published collections
**Appraise**
- DPL assess collection
- DPL confer with Digital Imaging Manager about what is on Drobo in relevant collection

**Plan**
- DPL inform DCL / IROL about what is in on Drobo for that collection
- DPL confer with DCL/IROL about what they want in DIPStore from collection
- Decide what data to keep

**Organize (DPL)**
- Delete files not being kept if interfering with reorganization
- Create Inventory Log
- Reorganize file structure
- Connect contents to presentation platform (if applicable)
- Assign UUIDs
- Rename Files
- Zip staging items and upload to staging

**Preservation**
- Create Processing log and transfer relevant info from Inventory Log
- Figure out Local ID format with relevant parties
- Package Items and update Processing Log
- Upload to APTrust
- Transfer to Synology

**Documentation**
- DPL add Collection Object Structure and Local ID decision to Confluence
- DCL or IROL add UUIDs to published collections

**Definitions**
- DPL = Digital Preservation Librarian
- DCL = Digital Curation Librarian
- IROL = Institutional Repository Outreach Librarian
Appraisal

DPL confer with Digital Imaging Manager about what is on Drobo in relevant collection

<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000_james_baldwin_program</td>
<td>7/14/2009 12:20 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2001_contract_deneuve</td>
<td>7/14/2009 12:44 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2008_interview_shelby_foote</td>
<td>7/14/2009 3:59 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2009_foote_greenbriar</td>
<td>7/14/2009 4:35 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2010_conversations_with_narrative_artist</td>
<td>10/8/2009 4:17 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2012</td>
<td>7/15/2009 4:55 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>2013_interview_shelby_foote</td>
<td>7/16/2009 1:14 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>books</td>
<td>7/16/2009 1:22 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Individual Letters</td>
<td>7/31/2009 10:00 AM</td>
<td>File folder</td>
</tr>
<tr>
<td>letters</td>
<td>10/13/2009 3:25 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>letters_partll</td>
<td>8/19/2009 4:32 PM</td>
<td>File folder</td>
</tr>
</tbody>
</table>
**Appraise**

- DPL assess collection

**Plan**

- DPL confer with Digital Imaging Manager about what is on Drobo in relevant collection
- DPL inform DCL / IROL about what is in on Drobo for that collection
- DPL confer with DCL/IROL about what they want in DIPStore from collection

**Organize (DPL)**

- Delete files not being kept if interfering with reorganization
- Create Inventory Log
- Reorganize file structure
- Connect contents to presentation platform (if applicable)
- Assign UUIDs
- Rename Files
- Zip staging items and upload to staging

**Preservation**

- Create Processing log and transfer relevant info from Inventory Log
- Figure out Local ID format with relevant parties
- Package Items and update Processing Log
- Upload to APTrust
- Transfer to Synology

**Documentation**

- DPL add Collection Object Structure and Local ID decision to Confluence
- DCL or IROL add UUIDs to published collections

---

DPL = Digital Preservation Librarian
DCL = Digital Curation Librarian
IROL = Institutional Repository Outreach Librarian
Planning

DPL = Digital Preservation Librarian
DCL = Digital Curation Librarian
IROL = Institutional Repository Outreach Librarian

- DPL inform DCL/IROL about what is in on Drobo for that collection
- DPL confer with DCL/IROL about what they want in DIPStore from collection
- Decide what data to keep

Table:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>jpg</td>
<td>10/21/2009 6:22 PM</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>pdf</td>
<td>11/6/2009 7:04 PM</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>tif</td>
<td>10/21/2009 6:20 PM</td>
<td>File folder</td>
<td></td>
</tr>
<tr>
<td>Letters</td>
<td>7/27/2009 9:21 AM</td>
<td>Microsoft Edge</td>
<td>47,178 KB</td>
</tr>
</tbody>
</table>
DPL = Digital Preservation Librarian
DCL = Digital Curation Librarian
IROL = Institutional Repository Outreach Librarian
Organization

• Before using python restructuring scripts, every collection needed to be organized in the following way:
  • Each object in its own folder
    • Object can include a variety of file types
    • More leeway than lab generated objects re: keeping access copies
  • Inventory Log generated for documentation
    • Helps Digital Curation Librarian add UUIDs to access platform records
Organization

PDF folder of library namesake collection

Tiff folder of library namesake collection
Automated Organization with Python

• 1_FolderInventoryLoader.py
  • User points script to a directory and it generates a CSV with folder paths in first column, folder name in second column, and an empty UUID column in 3rd column
  • User inputs UUIDs and saves CSV

<table>
<thead>
<tr>
<th>Directory Path</th>
<th>Folder Name</th>
<th>UUID</th>
</tr>
</thead>
<tbody>
<tr>
<td>C:\Users\gmurphy\D 19970106v21n11</td>
<td></td>
<td>UABP_REP000145</td>
</tr>
<tr>
<td>C:\Users\gmurphy\D 19970113v21n12</td>
<td></td>
<td>UABP_REP000146</td>
</tr>
<tr>
<td>C:\Users\gmurphy\D 19970120v21n13</td>
<td></td>
<td>UABP_REP000147</td>
</tr>
<tr>
<td>C:\Users\gmurphy\D 19970127v21n14</td>
<td></td>
<td>UABP_REP000148</td>
</tr>
</tbody>
</table>
Automated Organization with Python

- 2_FolderAndFileRenamer.py
  - Takes CSV generated by first script and it renames folders and files according to UUIDs
Automated Organization with Python

• 3_zipbs.py
  • Zips contents of folders and transfers them to DIPStore on Synology NAS
  • User inputs source directory and target directory
  • Able to tell script to ignore certain file extensions if Digital Curation Librarian did not want them in DIPStore
Documentation

Item Level Structure

/ARCH_WCL000001

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCH_WCL000001_0001a.pdf</td>
<td>PDF of a full letter</td>
</tr>
<tr>
<td>ARCH_WCL000001_0002a.tif</td>
<td>A tif image of one page of the letter</td>
</tr>
<tr>
<td>ARCH_WCL000001_0002a.txt</td>
<td>A transcript of the letter</td>
</tr>
</tbody>
</table>

Local ID

The Local ID for the letters is the UAB Archives collection number (MC122) followed by the original holding location, “MHSL” and accession number.

- Examples:
  - MC122_MHSL 1000

Exceptions

- None
Project Plan

- Process one digital collection at a time
  - Processed one collection to get an idea of timing, then appraised all other collections and estimated timing
- Develop a weekly plan (gantt chart)
- Process collections slated for Alma Digital first due to migration
  - Collections slated for Institutional Repository done in second half of project
## Project Plan

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nightingale Letters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proust</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical Collections Exhibits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pappas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phoenix Yearbook</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M.H. Sterne Collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>William Carter Collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Civil War</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reynolds Lectures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osler Letters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBHM Newsletters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StoryCorps Oral Histories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAB 50th Oral Histories and Podcast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral History Office</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portraits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Digital Commons</th>
<th>9-Jan</th>
<th>16-Jan</th>
<th>23-Jan</th>
<th>30-Jan</th>
<th>6-Feb</th>
<th>13-Feb</th>
<th>20-Feb</th>
<th>27-Feb</th>
<th>6-Mar</th>
</tr>
</thead>
<tbody>
<tr>
<td>eReporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty Works</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NELLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GreenMail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aura</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnographic Films</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAB Report(er)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaleidoscope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lessons Learned

- Your work has ramifications for years to come
- Plan heavily now, benefit later
Thank you!