A Collaborative Approach to Digital Preservation for the Five Colleges

Aaron Rubinstein
University and Digital Archivist
Special Collections and University Archives
University of Massachusetts Amherst

Shaun Trujillo
Digital Collections and Metadata Lead
Digital Assets and Preservation Services
Mount Holyoke College
The Five Colleges

Amherst College
Hampshire College
Mount Holyoke College
Smith College
UMass Amherst

Founded in 1965

Strong collaborative infrastructure

Digital resource collaboration new and experimental
Digital Preservation at the Five Colleges

- Digital Preservation Task Force formed in 2011
- First phase: introspection, self assessment, and research

Lesson learned:
Unless all institutions commit to a similar level of readiness, collaboration is impossible.
Three-Pronged Plan

- **Education**
  Digital Preservation Management Workshop
  POWRR workshop
  Readiness guide *

- **Best Practices/Standardization**
  Stakeholders and decision making

- **Experimentation**
  Archivematica pilot project

*https://www.fivecolleges.edu/libraries/digital-preservation/digital-preservation-a-guide-for-the-five-colleges
Enter Archivematica

- Micro-Service model of DP
- Excels at born-digital accessioning
- Customizable workflow
- Runs on Ubuntu Linux OS
- Two-part architecture:
  - Client (Pipeline)
  - Storage Service
Centralized Storage Service
- Server hosted at MHC (spike)

Pipelines - Local Clients running on VirtualBox virtual machine emulation (or not, physical Ubuntu machine)
- Clients connect to spike via VPN
  - reduces complication of two-way SSH traffic and VM network configuration
  - use NAT connection and sign in over VPN (no bridging, no port forwarding)

Project Leads administer the Storage Service
- gain experience assigning and administering transfer and storage of AIPs & DIPs, i.e. spaces and locations

Working Group collaborates on policies and use case workflows for their respective institutions. Configures local client to reflect those decisions.
Benefits of an Archivematica Pilot

- Applied Five College collaboration
- Cross Committee Working Group
- Jumpstart digital preservation conversations and decision making by focusing on something tangible
- Uncover and learn about implicit practices at the Five Colleges
  - Articulate practices in place
  - Align practices with policy/requirements
  - Define policy where there is none
  - Define content streams
- Create a ‘baseline’ for digital preservation in the Five Colleges
## Processing configuration

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Send transfer to quarantine</td>
<td>No</td>
</tr>
<tr>
<td>Approve normalization</td>
<td>Yes</td>
</tr>
<tr>
<td>Store AIP</td>
<td>Yes</td>
</tr>
<tr>
<td>Remove from quarantine after</td>
<td></td>
</tr>
<tr>
<td>Create SIP(s)</td>
<td>--Actions--</td>
</tr>
<tr>
<td>Extract packages</td>
<td>Extract</td>
</tr>
<tr>
<td>Normalize</td>
<td>--Actions--</td>
</tr>
<tr>
<td>Reminder: add metadata if desired</td>
<td>Continue</td>
</tr>
<tr>
<td>Select file format identification command (Transfer)</td>
<td>Fido version 1 PUID runs Identify</td>
</tr>
<tr>
<td>Select file format identification command (Ingest)</td>
<td>Use existing data</td>
</tr>
<tr>
<td>Delete packages after extraction</td>
<td>Delete</td>
</tr>
<tr>
<td>Select compression algorithm</td>
<td>7z using bzip2</td>
</tr>
<tr>
<td>Select compression level</td>
<td>5 - normal compression mode</td>
</tr>
<tr>
<td>Store AIP location</td>
<td>--Actions--</td>
</tr>
</tbody>
</table>

[Save button]
<table>
<thead>
<tr>
<th>Preservation</th>
<th>Nikon Digital SLR Camera Raw Image File</th>
<th>Transcoding to tif with convert</th>
<th>0 out of 0</th>
<th>View</th>
<th>Replace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preservation</td>
<td>Kodak Digital Camera Raw Image File</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Sony RAW Image File</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Canon CR2 Raw</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Raw Bitmap</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Generic TGA</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Truevision TGA Bitmap 2.0</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Truevision TGA Bitmap 1.0</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Generic PICT</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>PICT 1.0</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Adobe Photoshop</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Exchangeable Image File Format (Uncompressed) 2.2</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Exchangeable Image File Format (Uncompressed) 2.1</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Preservation</td>
<td>Exchangeable Image File Format (Uncompressed) 1.0</td>
<td>Transcoding to tif with convert</td>
<td>0 out of 0</td>
<td>View</td>
<td>Replace</td>
</tr>
<tr>
<td>Transfer</td>
<td>UUID</td>
<td>Transfer start time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample series</td>
<td>89a46845-0bcd-4917-a482-ea004a798b9a</td>
<td>2013-10-10 13:06</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Micro-service: Create SIP from Transfer**

- Job: Create SIP(s) ![?)
- Job: Load options to create SIPs
- Job: Check transfer directory for objects
- Micro-service: Complete transfer
- Micro-service: Characterize and extract metadata

**Actions**

- Create SIP(s) manually
- Send to backlog
- Reject transfer
- Create single SIP and continue processing

**Jobs**

- Load labels from metadata/file_labels.csv: Completed successfully
- Characterize and extract metadata: Completed successfully
- Identify file format: Completed successfully
- Determine which files to identify: Completed successfully
- Select file format identification command: Completed successfully
- Move to select file ID tool: Completed successfully
- Micro-service: Clean up names: Completed successfully
- Micro-service: Scan for viruses: Completed successfully
Micro-Services Inform Decision Making

- **Characterization**: managing a panoply of file extensions
  - Which formats are common? Which are edge cases?
- **Normalization**: Master file format / access file format
  - Generalized file management / discreet file management
  - Legacy formats >> Data Loss via normalization
    - Acceptable data loss vs. critical characteristics
- **Versioning** - Master, Access 1, Access 2, etc. LOCKSS
- **Metadata** compliance - at the object level, folder level, item level?
- **Custom Actions**: plugin scripts for specific use cases
  - e.g. Exif metadata extraction with ExifTool
  - e.g. provide OCR for PDFs with Tesseract
Questions?