Appraising Digital Archives with Archivematica

Michael Shallcross
University of Michigan Bentley Historical Library
shallcro@umich.edu
IEEE Big Data 2016: Computational Archival Science Workshop
December 8, 2016
Overview

• Background
  – Bentley Historical Library
  – 2014-2016 Grant Project
• Highlights of Archivematica’s Appraisal Functionality
• Scalability
• Conclusion: Next Steps and Related Projects
Bentley Historical Library

- [http://bentley.umich.edu/](http://bentley.umich.edu/)

- Established in 1935 as:
  - Official archives of the University of Michigan
  - Repository for Michigan individuals and organizations

- 2014 reorganization: consolidated physical/digital curation units

- Currently holds 25+ TBs of digital archives (born-digital, digitized, web archives)
Appraisal at the Bentley

• An iterative process.
• Similar goals for the appraisal of both physical and digital archives:
  – Understand the nature and scope of material.
  – Establish intellectual control to facilitate the arrangement and description of content.
  – Determine what/how much should be retained in permanent collections (resource management).
  – Identify potential preservation and rights issues (risk management).
  – Discover and verify the presence of sensitive personally identifiable information.
2014-2016 Grant Project

• Funded by Andrew W. Mellon Foundation
• “ArchivesSpace-Archivematica-DSpace Workflow Integration”
  – ArchivesSpace: open-source archival management software; create archival description and track locations.
  – Archivematica: open-source digital preservation system; prepares Archival Information Packages (AIPs).
  – DSpace: open-source repository platform; enables both preservation of and access to content.

• Project blog: http://archival-integration.blogspot.com/
Grant Project Goals

• Remove barriers and simplify tools for archivists to work with digital content
  – Facilitate the creation/reuse of metadata across systems.
  – Streamline the ingest and deposit of content in a preservation repository.

• Ensure that appraisal is a vital component of digital archives workflows.

• Find solutions that meet Bentley needs but are flexible and scalable for others.
Archivematica Appraisal Tab
Backlog Pane:
Directory Structure Review

- **View:**
  - Directory structure and hierarchy
  - Naming conventions
  - Overall organization

- Provides context related to:
  - Digital content
  - Records management practices
### Analysis Pane: File Format Characterization

<table>
<thead>
<tr>
<th>Format</th>
<th>PUID</th>
<th>Group</th>
<th>Number of files</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrobat PDF</td>
<td>fmt/19</td>
<td>Portable Document Format</td>
<td>3 objects</td>
<td>0.145 MB</td>
</tr>
<tr>
<td>Comma Separated Values</td>
<td>x-fmt/18</td>
<td>Text (Plain)</td>
<td>1 object</td>
<td>0.002 MB</td>
</tr>
<tr>
<td>Excel 97 Workbook</td>
<td>fmt/61</td>
<td>Spreadsheet</td>
<td>33 objects</td>
<td>22.011 MB</td>
</tr>
<tr>
<td>Excel for Windows 2007+</td>
<td>fmt/214</td>
<td>Spreadsheet</td>
<td>1 object</td>
<td>0.012 MB</td>
</tr>
</tbody>
</table>

**Charts:**
- Formats (by total number of files)
- Formats (by total size of files)
Analysis Pane:
Sensitive Data Identification

- “Examine Contents” tab
- Review bulk_extractor reports to ID sensitive personal information (SSN and CC’s).
- View list of files and contextual snippets (with ability to preview/download) to verify
Analysis Pane: Content Preview

• Review files (or a sample) to grasp intellectual content.
• Uses available browser plug-ins to display content.
• May also download files to render with local tools.
Tagging

- Apply tags to files and/or folders
- Add via:
  - Backlog Pane
  - File List Pane
  - Analysis Pane (“Examine Contents” tab)
- Serve as aide-mémoire:
  - Intellectual arrangement
  - Sensitive data
  - Deaccession decisions
  - Further review
Scalability

• Primary use case for Appraisal Tab development:
  – Digital archives of former U.S. Senator Carl Levin
  – 1.2 TB / 217,135 files

• Archivematica: limit of 80-90,000 files per transfer (due to performance issues).

• Solutions:
  – Short-term: modify workflows to chunk content
  – Long-term: improvements to Archivematica
Conclusion

• Grant completed on Oct. 31, 2016; still implementing locally.
• Appraisal Tab to be released in Archivematica v1.6 (2017)
• Micro-service design permits inclusion of additional tools:
  – Treemap visualizations
  – Brunnehilde integration
  – Named Entity Recognition (NER), Natural Language Processing (NLP), topic modeling
  – Hand-offs with related projects:
    • BitCurator (digital forensics platform)
    • ePADD (email curation platform)