

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/>
- 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

archivematica Transfer ⁶ Backlog Appraisal Ingest ⁸ Archival storage Preservation planning Access Administration bentley ▾

MMDP Any Keyword Search transfer backlog

Add New

Tags All

Backlog Analysis File list ArchivesSpace Arrangement

Backlog 0 transfers, 13 objects selected

Add tag to selected files

Deselect all Collapse all Expand all

- MMDP-c16c76b0-9240-4435-b3db-c6a1aed21938
- MMDP-finalAnswer-81c1b66f-0982-4ec1-8e1e-f943f98b0b6d
- MMDP-finalAnswer_2-8cdab1f7-e0b2-4673-91af-299ac3194e68
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b
 - objects
 - Club-Report.doc
 - FRPEnForm.pdf
 - MS-OfficeOpenXML-samples.zip-2016-10-14T13_45_50.817424_00_00
 - sampledocx.docx
 - samplepptx.pptx
 - samplexlsx.xlsx
 - Project.zip-2016-10-14T13_45_50.817424_00_00
 - Members_Master2009.xls
 - PPT_test.ppt
 - Syllabus_FINAL.doc
 - _datavibe-l_FW_job_vacancy.rtf
 - article.pdf

Objects Tags Examine contents Preview file

Report Visualizations

Formats (by total number of files)

Formats (by total size of files)

- Acrobat
- PDF (fmt/18)
- Excel 97
- Workbook (fmt/61)
- Excel for Windows 2007+ (fmt/214)
- Microsoft Word
- Document 97-2003 (fmt/40)
- Microsoft Word for Windows 2007+ (fmt/412)

File List 10 objects in list, 0 objects selected

Date range start

Date range end

Add tag to selected files

- Filename
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b/objects/_datavibe-l_FW_job_vacancy.rtf
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b/objects/article.pdf
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b/objects/Club-Report.doc
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b/objects/FRPEnForm.pdf
- MMDP-finalAnswer_3-70181c7c-e82a-4088-ae78-e15ed3bb006b/objects/Project.zip-2016-10-

ArchivesSpace

fire Identifier

Search ArchivesSpace

Add New Child Record Add New Digital Object

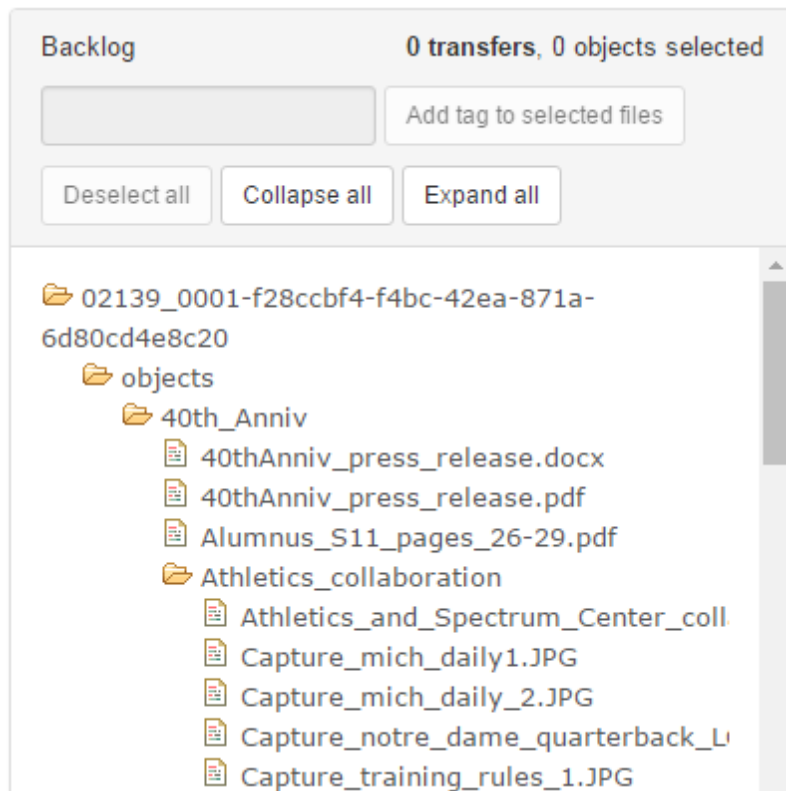
Edit Rights Metadata Edit Metadata Delete selected

Finalize Arrangement

▼ Ann Arbor (Mich.) Fire Department records (861042 Cc 2;CC Out. Vol.)

- Minute Books
 - Fire Department minute book, 1854-1888
 - Defiance Hood and Ladder Company (volunteer fire company)
 - Eagle Fire Department (volunteer fire company) minute book, 1850-1864
 - Mayflower Fire Engine Company minute book, 1866-1869
 - Administrative Records
 - Poor Relief

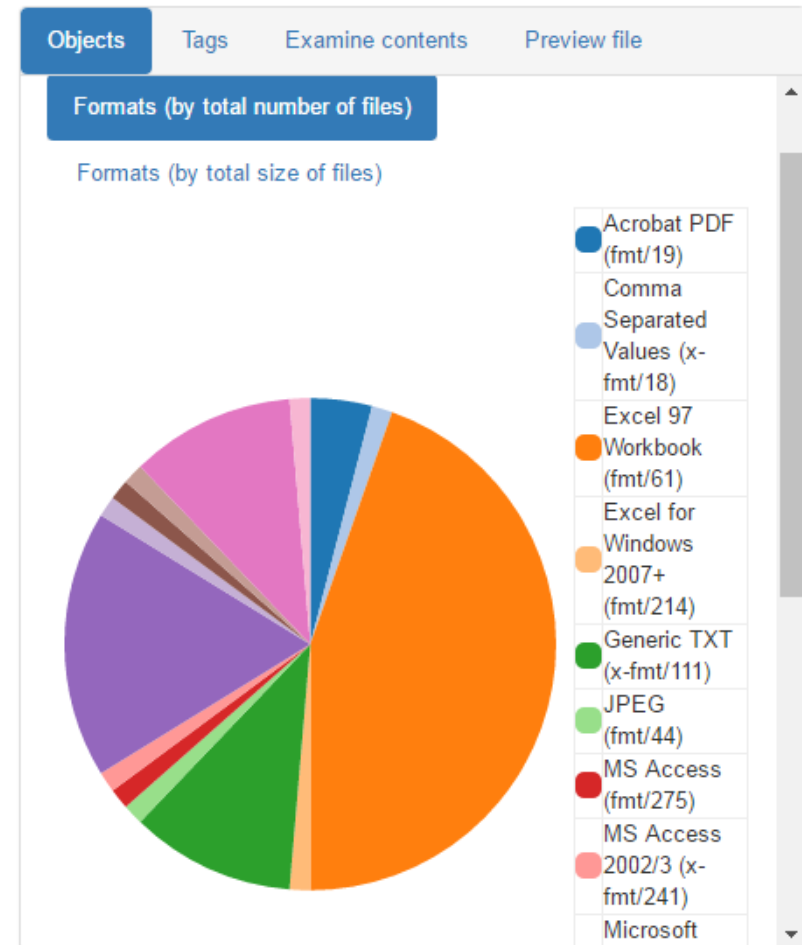
Backlog Pane: Directory Structure Review



- View:
 - Directory structure and hierarchy
 - Naming conventions
 - Overall organization
- Provides context related to:
 - Digital content
 - Records management practices

Analysis Pane: File Format Characterization

Objects				
Objects	Tags	Examine contents	Preview file	
Report		Visualizations		
Format	PUID	Group	Number of files	Size
Acrobat PDF	fmt/19 ↗	Portable Document Format	3 objects	0.145 MB
Comma Separated Values	x-fmt/18 ↗	Text (Plain)	1 object	0.002 MB
Excel 97 Workbook	fmt/61 ↗	Spreadsheet	33 objects	22.011 MB
Excel for Windows 2007+	fmt/214 ↗	Spreadsheet	1 object	0.012 MB



Analysis Pane: Sensitive Data Identification

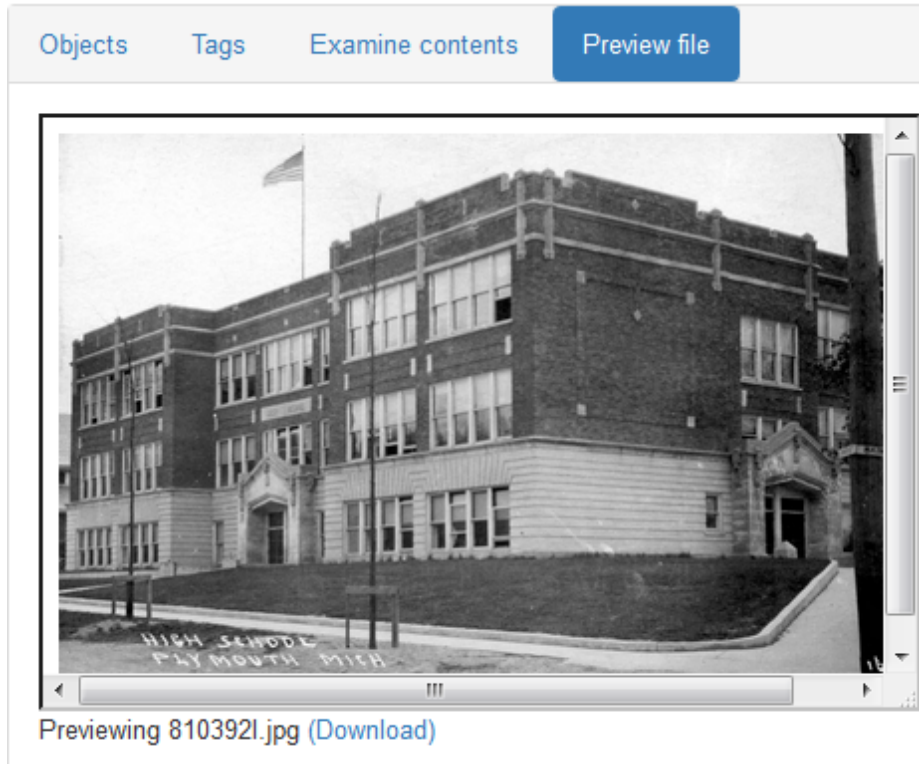
Files

		Add tag to checked files	
<input type="checkbox"/>	Filename	Preview	Bulk Extractor Logs
<input type="checkbox"/>	Contacts.pptx	Preview	
<input type="checkbox"/>	Tax_Return_2008.pdf	Preview	
<input type="checkbox"/>	request.txt	Preview	

Content	Context
SSN: 629369510	202 520-38-7202 SSN: 629369510 SSN: 262698143
SSN: 262698143	SSN: 629369510 SSN: 262698143

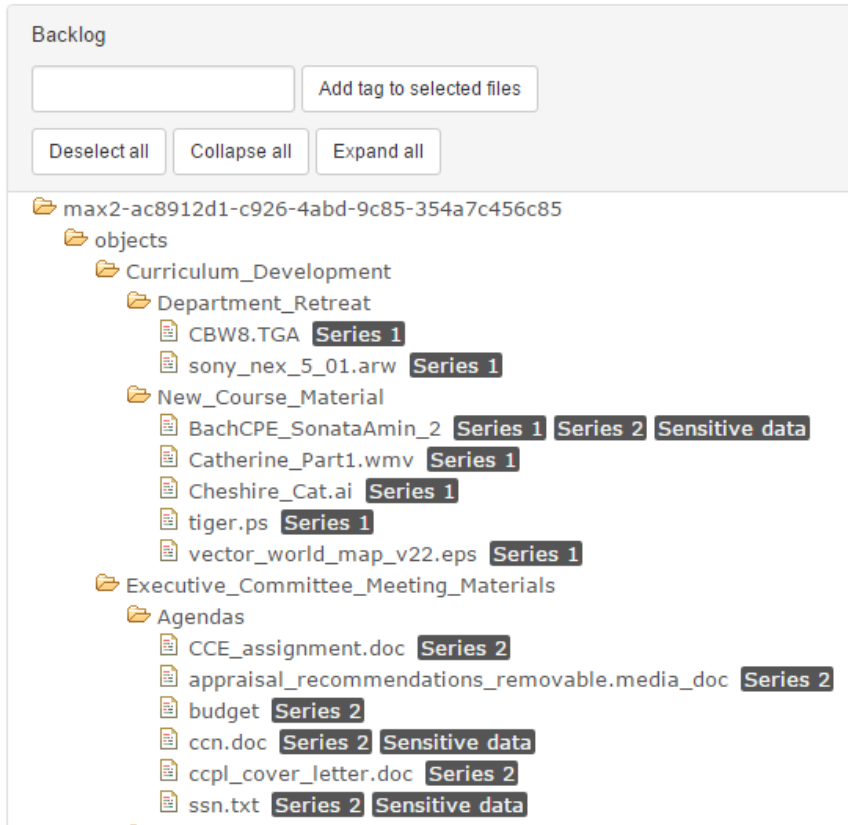
- “Examine Contents” tab
- Review [bulk extractor](#) reports to ID sensitive personal information (SSN and CC’s).
- View list of files and contextual snippets (with ability to preview/download) to verify

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
- Archivemata: limit of 80-90,000 files per transfer (due to performance issues).
- Solutions:
 - Short-term: modify workflows to chunk content
 - Long-term: improvements to Archivemata



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)

