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
UNIVERSITY OF MICHIGAN

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)

BENTLEY
HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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.

BENTLEY
HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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/

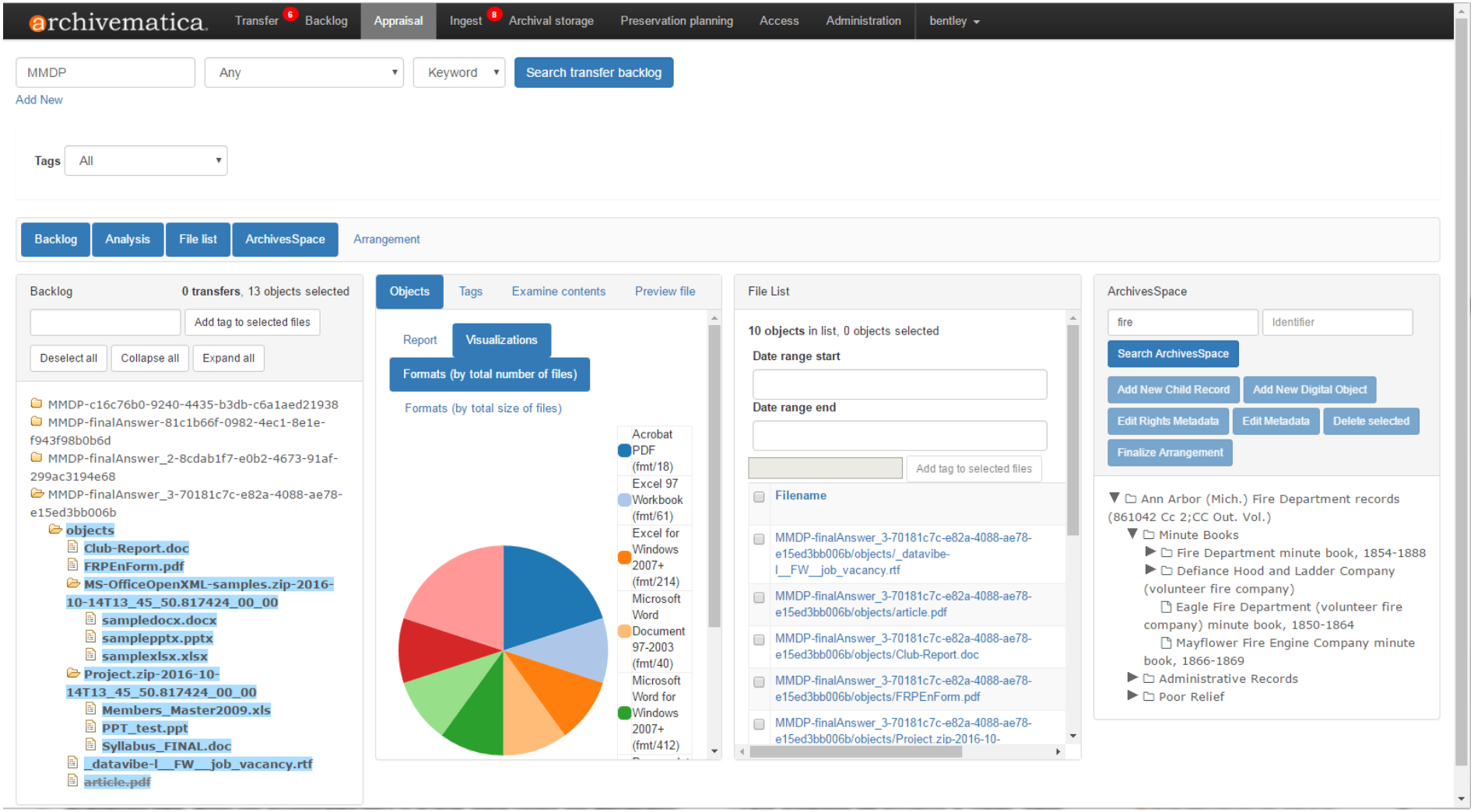
BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

Grant Project Goals

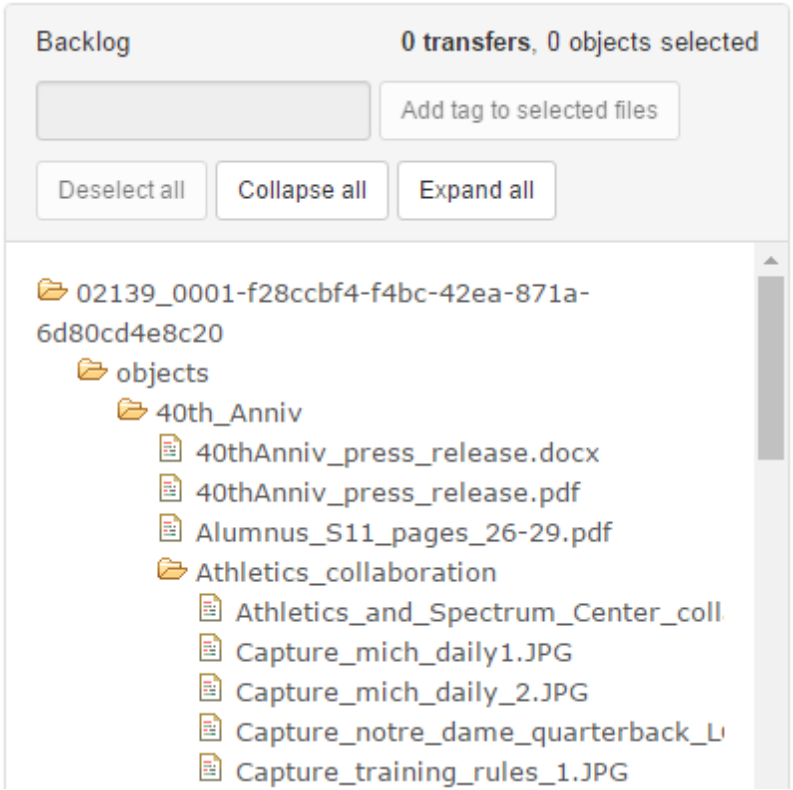
- Remove barriers and simplify tools for archivists to work with digital content
  - Facilitate the creation/reuse of metadata across systems.
  - Streamline the ingest and deposit of content in a preservation repository.
- Ensure that appraisal is a vital component of digital archives workflows.
- Find solutions that meet Bentley needs but are flexible and scalable for others.

BENTLEY
HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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

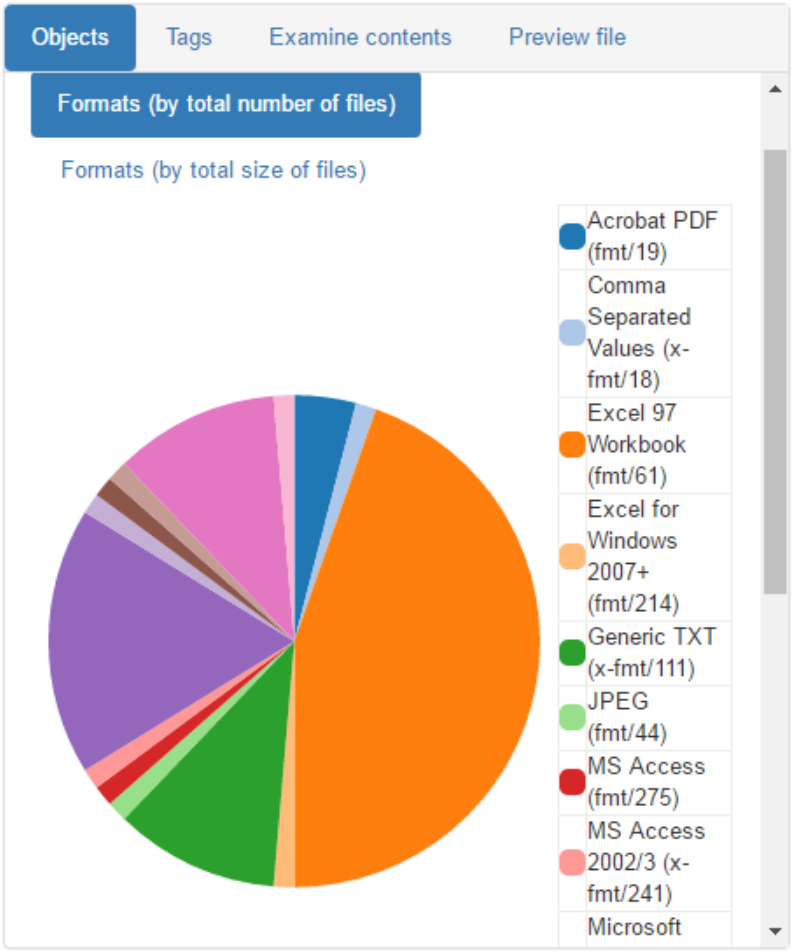
BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

Analysis Pane:
File Format Characterization

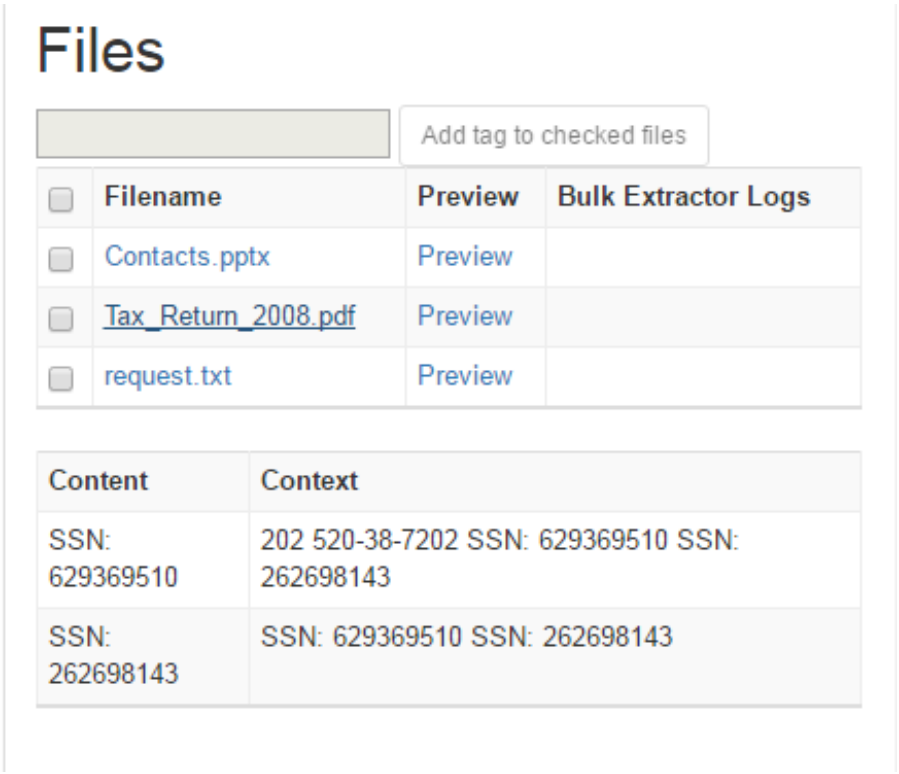
Objects | Tags | Examine contents | Preview file

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

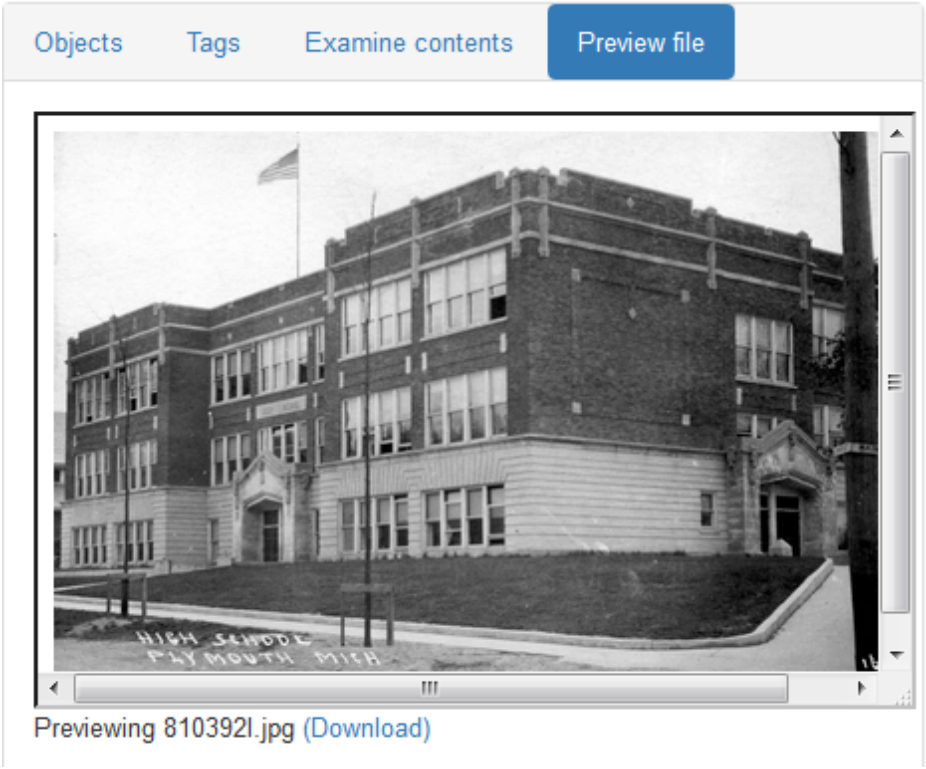
Filename	Preview	Bulk Extractor Logs
Contacts.pptx	Preview	
Tax_Return_2008.pdf	Preview	
request.txt	Preview	

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

BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

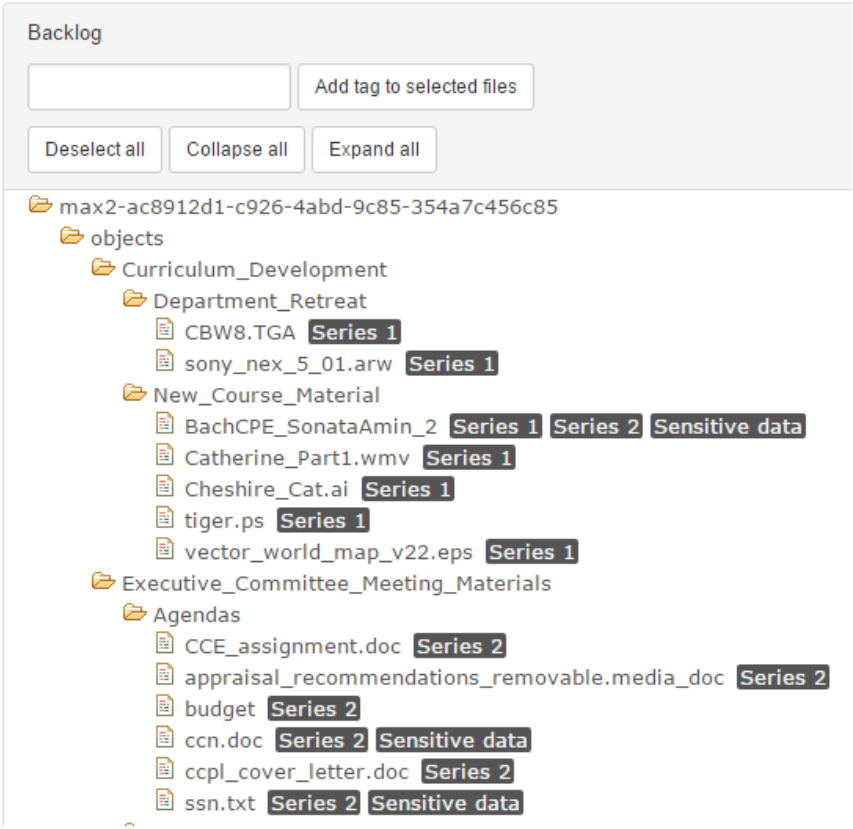
Analysis Pane: Content Preview



- 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.

BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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

BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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

BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN

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)

BENTLEY HISTORICAL LIBRARY
UNIVERSITY OF MICHIGAN