Appraising Digital Archives with Archivematica

Michael Shallcross
Bentley Historical Library
University of Michigan
Ann Arbor, MI (U.S.A.)
shallcro@umich.edu

Abstract—**The Bentley Historical Library, funded by a generous grant from the Andrew W. Mellon Foundation, has developed a new Appraisal and Arrangement tab in the Archivematica digital preservation system as part of its "ArchivesSpace-Archivematica-DSpace Workflow Integration" project. This new functionality permits users to conduct large-scale appraisal of digital archives as part of an end-to-end digital archives workflow.**

Keywords-appraisal, digital archives, digital preservation, open source software

I. Introduction

Established in 1935 by the University of Michigan Regents, the Bentley Historical Library ("the Bentley") serves as the official archives of the university and documents the history of the state of Michigan and the activities of its people, organizations and voluntary associations. In April 2014, the Bentley partnered with the main University of Michigan Library to launch a two-year project (funded by a generous grant from the Andrew W. Mellon Foundation, with a completion date of October 31, 2016) to integrate the functionality of three open source platforms (ArchivesSpace, Archivematica, and DSpace) in an end-to-end digital archives workflow. In doing so, archivists at the Bentley Historical Library sought to expedite the ingest, description, and preservation of digital archives by facilitating the creation and reuse of descriptive and administrative metadata among the above-mentioned open-source platforms; introducing features to appraise and characterize materials; and streamlining the deposit of fully processed content into a digital preservation repository. This paper will provide information on the Bentley's approach to the ingest and appraisal of digital archives, introduce the grant funded project, and highlight key project outcomes.

II. Appraisal and Digital Archives at the Bentley

In working with digital archives, the Bentley applies the same archival principles that inform our handling of physical collections, with added steps to ensure the authenticity, integrity, and security of content. As a result, "appraisal" tends to be an iterative process that often begins during initial conversations with donors as field archivists identify materials of interest to the archives and then recurs during the actual selection of content for transfer, the accession process, and the arrangement and description of materials by processing archivists.

In each iteration, the goals of appraisal are broadly the same: archivists seek to understand the intellectual content and scope of materials to determine if they should be retained as part of our permanent collections. Our curation staff and/or field archivists are sometimes able to review content (or a sample thereof) prior to its acquisition and accession, a process that helps us pinpoint the materials we are interested in and avoid the transfer of content that we have identified as out of scope or superfluous.

This pre-accession appraisal may not be possible for various reasons (technical issues, geographic distance, scheduling conflicts, etc.), but even then the record creators or donors are usually able to provide some level of understanding about the nature of the digital content and its relationship to our collecting policy, whether it is a high-level overview or item-level description in a spreadsheet.

Once content has been acquired and accessioned, appraisal is a crucial part of our ingest workflow that helps the archives to:

- Establish basic intellectual control of the content, directory structure, and/or original storage environment to facilitate the arrangement and description of content.
- Identify content that should be included in our permanent collections as well as superfluous or out-of-scope materials that will be separated (deaccessioned).
- Determine potential preservation issues posed by unique file formats, content dependencies, or other hardware/software issues.
- Address copyright or other intellectual property issues by applying appropriate access/use restrictions.
- Discover and verify the presence of sensitive personally identifiable information such as Social Security and credit card numbers.

These appraisal goals have informed the Bentley's procedures for digital archives since 2011 and have played an important role in the planning and development work conducted as part of our "ArchivesSpace-Archivematica-DSpace Workflow Integration" project.

III. Grant Project Background and Context

The Bentley has successfully managed and preserved large collections of electronic records since 1997, when former President James J. Duderstadt donated the contents of his personal computer to the archives. For much of this time, however, the library took a 'project' based approach to working with digital materials: unique processes were crafted for each new acquisition and procedures were highly manual. A 2010-2011 grant from the Mellon Foundation ("Email Preservation at the University of Michigan") provided the Bentley with sufficient staff and resources to develop more standardized workflows. Following the grant's conclusion, Michael Shallcross developed a local tool called the AutomatedProcessor (or "AutoPro"; see http://hdl.handle.net/2027.42/95923) that provided a command line interface to guide staff through a multistep workflow that involved more than twenty applications and utilities. While AutoPro ensured greater consistencies in procedures and standardized metadata collection and creation, the tool was never intended to be a long-term solution. The application employs a series of shell scripts (written by an archivist!) that have limited error handling, the component programs must be installed and frequently updated on individual workstations, and scalability becomes an issue, as very large files or large collections can require a large amount of workstation resources.

Faced with these challenges, the Bentley looked to emerging open-source archival tools to find a successor platform. The three platforms ultimately decided upon encompassed key functional requirements in the Bentley's approach to digital archives:

- ArchivesSpace (ASpace) is an open-source archival management software that combines the best features of Archon and Archivists' Toolkit. This system permits institutions to track accessions, manage collections, and generate Encoded Archival Description (EAD) finding aids and MARCXML. Development of ASpace was funded by the Andrew W. Mellon Foundation (2011-2013) and LYRASIS now serves as its institutional home.
- Archivematica is a free and open-source digital preservation system developed by Artefactual Systems (British Columbia). Archivematica employs a micro-service design to "provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model" and furthermore employs METS and PREMIS to record and track descriptive, administrative, and rights metadata.
- DSpace is an open-source repository platform that "preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets." Initially developed by MIT Libraries with partial support from a Mellon Foundation grant, DSpace has acquired a growing community of developers and is employed by the University of Michigan and approximately 1,400 other academic, nonprofit, and commercial organizations around the world.

As stated in the Bentley's proposal to the Mellon Foundation, ArchivesSpace will be employed to store descriptive, administrative, and rights metadata related to digital archives; Archivematica will be used to ingest content; associate it with descriptive metadata from ASpace, and prepare information packages for deposit; DSpace (hosted by the main University Library and used by the Bentley since 2008) will serve as a preservation repository and access portal for collections.

IV. Project Goals and Key Development tasks

The Bentley's formal goals for the project reflect our desire to increase internal capacity and at the same time contribute to the profession's growing interest in strategies and tools for working with digital archives. These core goals include:

- Facilitating the creation/reuse of metadata across systems.
- Streamlining the ingest and deposit of content in a preservation repository.
- Ensuring appraisal is a vital component of digital archives workflows.
- Finding solutions that meet Bentley needs but are flexible and scalable for others (so that an institution may adopt some, none, or all of the new features). This includes employing open standards and open-source technologies so that other institutions can modify or extend features to suit local needs.
- Sharing code and documentation with archives and digital preservation communities.

To accomplish these goals, the Bentley worked with project developers at Artefactual Systems to identify three key development tasks, as listed below.

A. Appraisal and Arrangement Tab

When the Bentley first started exploring the Archivematica digital preservation system, it was essentially a platform for running a defined series of preservation actions to transform a Submission Information Package into an Archival Information Package and a Dissemination Information Package. The interface included no features or functionality that would permit users to engage with or explore content—all such work was expected to occur outside of the Archivematica. The Bentley thus proposed the creation of a new functional space in the Archivematica dashboard—a new "Appraisal and Arrangement" tab (see https://wiki.archivematica.org/Appraisal_Arrangement_tab)—that would permit users to explore and characterize content, identify sensitive data, and preview files to understand the information therein. The tab is comprised of five "panes" that can be toggled on or off depending upon individual preferences and workflows and which include:

- Backlog pane: displays files that have been transferred into Archivematica and have undergone initial preservation actions (such as virus scans and checksum calculation)
- Analysis pane: includes features that facilitate the review and characterization of content (more below).
- File list pane: provides a listing of files within a given selection of content; is sortable by file path, file size, etc.
- ArchivesSpace pane: integrates with an instance of ArchivesSpace to facilitate the arrangement and description of digital content (more below).
- Arrangement pane: allows non-ArchivesSpace users to arrange digital content into appropriately-structured submission information packages.

B. *ArchivesSpace Integration*

As mentioned above, the Appraisal and Arrangement tab has a direct integration with ArchivesSpace. In conceiving of this new tab, the Bentley wanted to use the information yielded by the appraisal process to inform decisions regarding the intellectual arrangement and description of content. Rather than have multiple windows and applications open and running simultaneously, the Bentley suggested integrating basic functionality of the ArchivesSpace archival management system within Archivematica. This development resulted in a dedicated ArchivesSpace pane that uses the application's API to display the intellectual hierarchy of materials in a collection and permits users to add new archival objects to this arrangement, create and edit basic descriptive metadata and PREMIS rights statements, and drag/drop digital content from the backlog pane onto ASpace archival description, thereby effectively uniting data and metadata in a Submission Information Package.

C. *DSpace Integration*

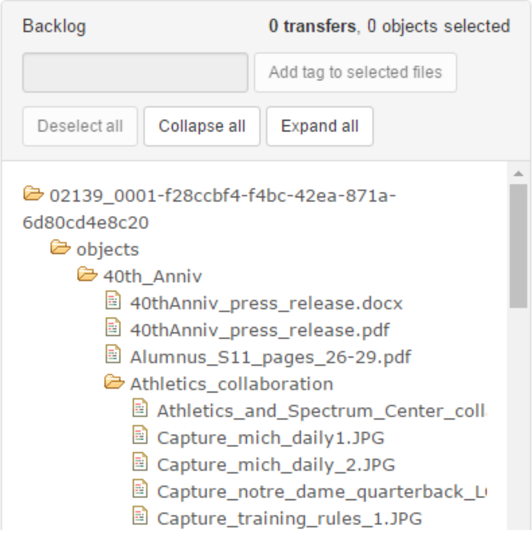
The final key development task in the project involves integration of the DSpace repository system with Archivematica and ArchivesSpace. Once arrangement and description is complete and a user has finalized a Submission Information Package in the ASpace pane, Archivematica will complete its Ingest workflow according to user specifications. Once this phase is complete, the resulting Archival Information Package (and/or Dissemination Information Package) will be automatically deposited to a DSpace collection via the SWORD protocol and DSpace REST API using information entered in the ASpace (or Arrangement) pane. SWORD will then return the DSpace handle so that Archivematica can update the location in its Storage Service and create a digital object record (with handle URI) in ASpace.

V. Highlights of Project Appraisal Functionality

Archivematica's new Appraisal and Arrangement tab facilitates the review and appraisal of content in a number of ways, listed below.

A. *Directory Structure*

First and foremost, the backlog pane will permit users to view the folder hierarchy, naming conventions, and overall structure of content in a given accession of digital archives (see Fig. 1). This information can provide important context about the digital content as well as the records management practices of the creators.



```
02139_0001-f28ccbf4-f4bc-42ea-871a-
6d80cd4e8c20
  objects
    40th_Anniv
      40thAnniv_press_release.docx
      40thAnniv_press_release.pdf
      Alumnus_S11_pages_26-29.pdf
      Athletics_collaboration
        Athletics_and_Spectrum_Center_coll
        Capture_mich_daily1.JPG
        Capture_mich_daily_2.JPG
        Capture_notre_dame_quarterback_L
        Capture_training_rules_1.JPG
```

Figure 1

By selecting a given folder in the Backlog pane, users can also review all associated files (including those in subfolders) within the File List pane, with additional options to sort based on file path/name as well as file size.

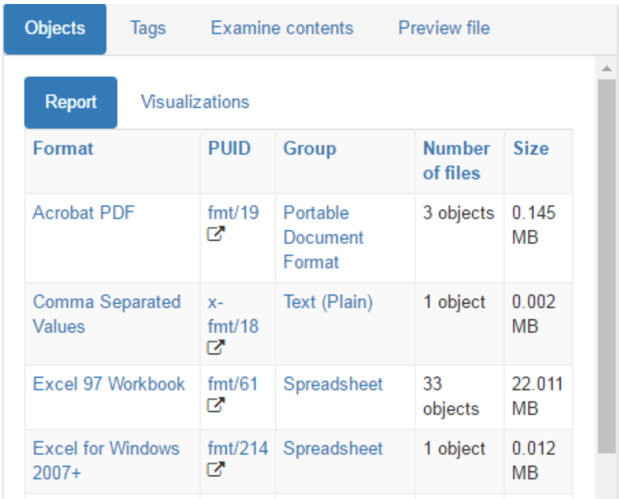
B. *File Format Characterization*

Viewing a distribution of file formats in an accession can be very helpful to better understand the range of materials and assist with preservation planning. File format characterization becomes that much more important in the event that high-value content is in a unique file format or is part of a complex digital object that requires additional preservation actions.

Archivematica allows users to choose among several options to identify and characterize file formats, including FIDO and Siegfried (both of which rely upon the PRONOM format registry). The Analysis pane in the new Appraisal and Arrangement tab presents this information in two forms. First, a table of file format information permits users to sort information about an accession by file format name, PRONOM Unique Identifier (PUID), file format group (i.e., 'word processing files', 'image files', etc., as defined within Archivematica's Format Policy Registry), and then the number of files and overall size on disk for content in a given format (see Fig. 2).

The Analysis pane also provides a visual representation of the file format distribution in a pie chart, with the option of viewing a representation of the format distribution by the total number of files or the total size of content on disk (see

Fig. 3). Hovering above different wedges of the chart will reveal the format name and PUID.



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

Figure 2

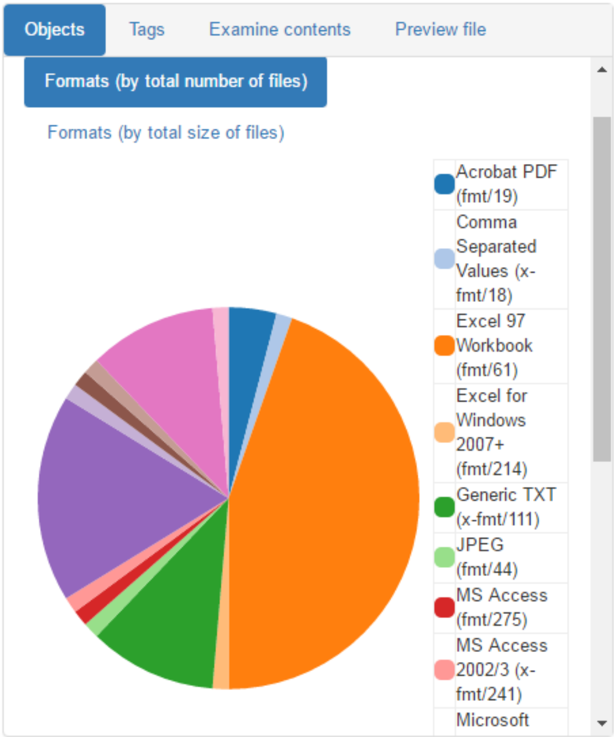
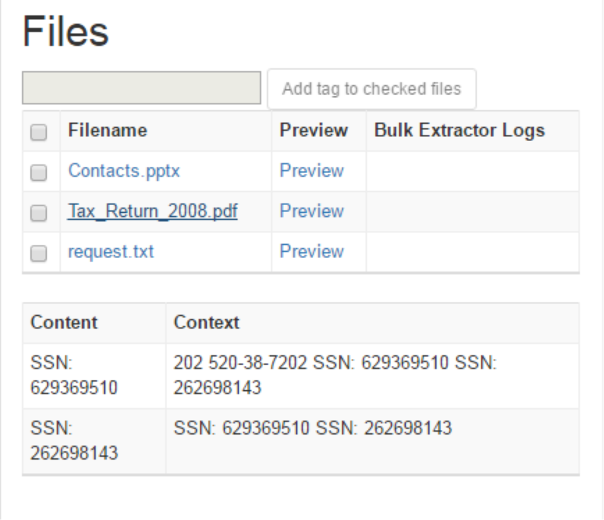


Figure 3

C. *Sensitive Data Identification*

Another important appraisal tool introduced as part of the Bentley's Mellon grant is the ability to review bulk_extractor log files to identify potentially sensitive personal information (specifically, Social Security and credit card numbers). Archivematica has employed the bulk_extractor tool, a resource developed for digital forensics activities, for some time but users previously had no way to review its output. Now, the "Examine Contents" section of the Analysis pane allows users to not only see which files were flagged as having potentially sensitive information, but also permits them to view a short snippet of text to verify if the information should be restricted or if it is merely a false positive (see Fig. 4).



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

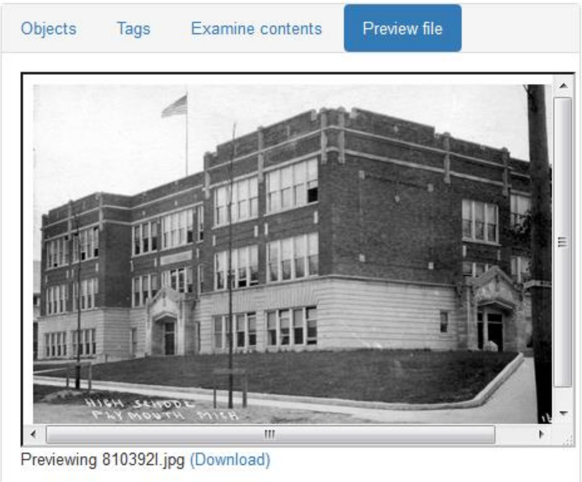
Content	Context
SSN: 629369510	202 520-38-7202 SSN: 629369510 SSN: 262698143
SSN: 262698143	SSN: 629369510 SSN: 262698143

Figure 4

If the scan did uncover truly sensitive data, the user can then apply a tag so that the content can be tracked during the appraisal and arrangement workflow (more on this below).

D. *Previewing Content*

While the appraisal measures previously identified in this paper are very helpful for understanding the nature and scope of a given accession of digital archives, it is also important for an archivists to review the files themselves (if only a sampling) to understand the intellectual content so that it can be described at an appropriate level of granularity. To this end, the Analysis pane also includes a "Preview" window (see Fig. 5).



Previewing 810392l.jpg (Download)

Figure 5

In its current incarnation, the Preview will display any file that can be opened via the user's browser. If a file is unable to be rendered from within the browser (or if the user needs to examine it more closely with specialized software), a copy may be downloaded locally.

E. Tagging

While engaging with content in the Appraisal and Arrangement tab, users are gleaning lots of information about the context, contents, and interrelations among the digital archives in a given accession. To help users keep track of this information (and to avoid having to rely on paper notes or aides memoires stored in another system), the Bentley introduced tagging functionality into the project. These 'tags' are intended to allow users to note anything of significance about entire directory structures or individual files—information such as possible intellectual arrangement, the presence of sensitive information, content that should be deaccessioned, or even just notes for later review. The tags are not persistent (they are stored within the system database while content is being reviewed in the Appraisal and Arrangement tab) and will not trigger any events within Archivematica's workflows. However, users can filter materials in the Backlog pane to only display items with specific tags, a feature which will aid in dragging and dropping items from the Backlog pane into the ArchivesSpace or Arrangement panes.

VI. Conclusion

The appraisal of digital archives is an important step for gaining intellectual control of materials, managing resources, and facilitating the use of collections. In developing the new Appraisal and Arrangement tab in the Archivematica digital preservation system, the Bentley Historical Library and its development partners, Artefactual Systems, sought to provide broadly-applicable tools and functionality that will assist users across the archives and digital preservation communities. As mentioned above, the ArchivesSpace-Archivematica-DSpace Workflow Integration project will conclude on October 31, 2016 and the project deliverables will be available in the 1.6 version of Archivematica, set to be released in the fall of 2016. Moving forward, the Bentley hopes that other institutions will extend this functionality and introduce new tools or integrations (for instance, managed hand-offs between the BitCurator suite of tools and Archivematica). For more information on the project (including elements beyond the scope of this paper), please see the project blog at: http://archival-integration.blogspot.com/

Acknowledgments

The Bentley Historical Library thanks the Andrew W. Mellon Foundation for its generous support of its ArchivesSpace-Archivematica-DSpace Workflow Integration project. Some elements of this paper originally appeared in the Bentley's project blog, located at: http://archival-integration.blogspot.com/