Born Digital Archives Working Group (BDAWG)
Email Archiving Task Force

Final Report

November 1, 2019

Task Force Members

Phase 1
Mary Caldera, Manuscripts and Archives (Chair)
Alice Prael, Beinecke Library
Jessica Quagliaroli, Manuscripts and Archives
Rachel Chatalbash, Yale Center for British Art
Matthew Gorham, Beinecke Library
David Cirella, Digital Preservation
Kevin Glick, Manuscripts and Archives
Michael Lotstein, Manuscripts and Archives
Jonathan Manton, Music Library

Phase 2 (report authors)
Jonathan Manton, Music Library (Chair)
David Cirella, Digital Preservation
Kevin Glick, Manuscripts and Archives
Michael Lotstein, Manuscripts and Archives

1. Summary

During FY19, the Born Digital Archives Working Group (BDAWG) formed an Email Archiving Task Force, comprised of current BDAWG members, plus Kevin Glick and Michael Lotstein from Manuscripts and Archives. The principal aim of this task force was to investigate how current tools and workflows in the field of email archiving might be implemented within Yale University Libraries and Museums (YUL/M), and integrated with existing tools and services used by YUL/M's special collections units, including the Digital Accessioning Support Service (DASS), Preservica and ArchivesSpace. The primary focus on this task force's work was the pre-acquisition, acquisition, accessioning and preservation of email collections. Collection description and discovery and access were considered mostly out of scope. However, through the work of this task force, information about these out of scope areas surfaced that was deemed useful, and was therefore documented and is in some cases noted in this report. For example, due to the ability of one of the tools we investigated, ePADD, which includes an access module, the

task force did consider how this module might be used to provide access to collections within a reading room environment. External remote access was not considered.

The task force's work incorporated:

- Identifying the stages/steps involved in acquiring, accessioning, and preserving email collections.
- Identifying general requirements for systems and software to support these stages/steps.
- Matching requirements with existing software currently being used within the archives community for email archiving.
- Evaluating software applications.
- Developing suggested workflows for acquiring, accessioning, preserving, and ultimately providing access to email collections within a reading room environment, using suggested software.

Initial work incorporated an environmental scan of units within YUL/M who are already collecting and working with email collections, as well as external organizations who are known to be actively working with email. This environmental scan helped us identify several tools actively being used within the archives community when working with email, as well as workflow considerations.

2. Environmental Scan

The environmental scan brought to light the shared use of several tools for email acquisition, appraisal, description, preservation, and access. Though appraisal, description and access were not the main focus of the scan as they were out of scope for the task force, they were two areas that came up consistently and so were considered and reviewed.

Common themes that surfaced were the need for iterative testing of tools and documentation of that testing; the need to scan for sensitive information; and consideration of future researcher access throughout the process of working with email collections. Additionally, the scan made it clear that preliminary workflows for email archiving work need to be flexible and adaptable as tools are tested and integrated into production use.

Below is a list of tools used for each stage of the email archiving process. The repositories either using or testing those tools are included in parenthesis.

Appraisal

- ePADD (Johns Hopkins, Stanford, MIT)

Acquisition

- ePADD (MIT)
- Aid4Mail for conversion (Northwestern, YCBA)
- Emailchemy for conversion (Duke, Canadian Centre for Architecture)
- readPST CLI tool in BitCurator Suite (Northwestern)
- Mail converted to MBOX (Northwestern, MIT, BRBL)
- Mail converted to .PST (Northwestern, MIT, Canadian Centre for Architecture, BRBL, YCBA)
- Mail converted to .EML (BRBL)

Accessioning

- ArchivesSpace (MIT)
- Archivematica (Johns Hopkins)

Description

- ePADD (Stanford)
- Siegfried/Brunnhilde (Duke)
- Archivematica (Johns Hopkins)

Preservation

- Bagger to create SIPs (Northwestern)
- Archivematica (Johns Hopkins, Canadian Centre for Architecture)
- MIT may ingest in the future

Access

- ePADD (Northwestern, Duke, CCA)
- QuickView Plus (BRBL)
- FTK (BRBL)

3. Core requirements with matching tools

Following on from the Environmental Scan, the task force created a set of core requirements for workflows and tools. These were identified over the course of a multi-stage process that included the creation of user stories and an analysis of the available tools, enabling the selection of software relevant to our needs.

a. User Stories

Early in the process the task force set out to determine workflow and tool requirements. This was achieved through generating User Stories that outlined actions that would need to be taken by key personas, from the point of reviewing

email collections to ingesting them into a digital preservation system. A total of 24 user stories were created for the following stages: collection development and appraisal, acquisition and accessioning, description, preservation, and collections management. The full text of our user stories are collected in Appendix B.

b. Identification of BDAWG requirements

Based on a discussion of the user stories, the task force identified thirty in-scope core requirements within the following categories: General System Requirements, Pre acquisition appraisal, Acquisition, Accessioning and Preservation. Though out of scope, the task force also identified requirements for Post acquisition appraisal and review, Description, and Access. The full text of these identified requirements are collected in Appendix A.

c. Overview of Tools

Drawn from the environmental scan and a search informed by the core requirements specified by the task force, a set of eleven relevant tools were evaluated through documentation review and preliminary testing, to identify viable tools to satisfy each of the identified requirements. See Appendix C for an overview of these tools.

d. Core requirements with matching tools

In order to identify the tools that the task force would investigate in more depth and formulate potential workflows for, those tools that offered relevant functionality were paired with each of the Core Requirements outlined in Appendix A.

The following core requirements had been previously identified:

General System Requirements

- Ability to take action on emails in bulk and individually
- Ability to track review process of a collection during appraisal / review process
- Access and use restrictions must be able to be implemented by the systems

Pre-acquisition Appraisal

- Review content prior to acquisition and reporting capability
- Select/deselect messages and folders for acquisition
- Ability to flag materials to be redacted or embargoed before accessioning.PII scan for messages that maybe should be redacted or not transferred as they include PII
- Able to select/deselect messages from chosen individuals

Acquisition

- Acquire email directly from email service API (IMAP and SMTP)
- Capture attachments and retain the relationship with messages
- Acquire email directly from email client in native export format
- Run virus scan on incoming messages and attachments
- Temp secure storage for content if not accessioned immediately
- Capture data and of what acquired including fixity
- Direct transfer to temporary secure storage location
- Option to include/exclude messages/folders based on selecting and flagging done by donor
- Retain tagging of messages/folders done by donor re: embargo, security

  Accessioning (and Arrangement & Description as far as it is part of accessioning)
    - Run fixity check on incoming messages and attachments
    - Metadata extraction from messages and attachments (e.g file formats)
- PII scan of text
- De dupe messages
- Export directly to Preservica
- Accession record creation in ASpace
- Documentation of mailbox structure (directory listing - including folder names) and list of correspondents
- Temp secure storage for content if not ingested into Preservation system immediately
- Extract metadata to create DACS-compliant single-level minimum description (i.e. collection title, inclusive dates; extent; name of creator(s); scope and contents; conditions governing access; languages). Of course some of this info could be acquired at the point of acquisition through communication with the donor/seller.
- report on summary data, eg number, correndents, etc
- Apply restrictions to embargo items / assign disposition dates
- Run virus scan on incoming messages and attachments
- Create normalized derivative in mbox or pst

Preservation

- Email format must be accepted by Yale's Digital Preservation System (Preservica) - this may require a normalization step (to EML, PST or MBOX) for email exported in an unaccepted format
- Normalize attachments as required
- Maintain relationship between messages / folders contained within a single SIP

The applications Aid4Mail, ePADD, and FTK were determined to satisfy the majority of these requirements, and so were therefore selected for further investigation.[1]

4. **Tools evaluation and workflow development**

Following the work outlined in section 3, a small working group was charged to further evaluate tools that met the defined core requirements. This work focused primarily on ePADD. Initially, the working group struggled in their evaluation process as different applications had sometimes significantly different functionality and because no set testing/evaluation criteria had been established or could easily be found in a literature search. In lieu of the complete in-depth evaluation of all software applications, the group focused on whether and how systems already in use somewhere at Yale in addition to ePadd (which was specifically requested by the task force to be included) could be utilized to fulfil the core requirements. After some testing of applications, the working group eventually decided to develop process workflow diagrams that would model how a staff member might undertake the processes of pre-acquisition, acquisition, accessioning, and providing access to email in a way that adheres to the core requirements.

The workflow diagrams are not proscriptive of a single procedure or software tool, but are cognizant of slightly different policies and collecting scenarios in the different Yale special collections. The pre-acquisition process would only be undertaken in some situations where donors/sellers are available to undertake the work. The group determined ePADD was the only application that would be utilized in this process. The acquisition process was determined to be able to be accomplished using either the ePADD processing module, or Forensic Toolkit (FTK), or Aid4Mail, with different branches of the model tracing the workflow path for each. The email accessioning process was determined to be accomplished using a combination of FTK or Aid4Mail along with Preservica and some custom programming. The activity of providing access to email was scoped to within secure Yale special collections reading rooms, using the computer configuration proposed in YUL/M's base image project, using either the ePADD Delivery Module with temporarily saved access copies, or email saved in any number of different formats exported from Preservica to proprietary email applications running on the base image locked computers.

With these diagrams, it was hoped that staff could begin to build out component parts of the process iteratively and test each with real-world examples. However, in practice,

[1] Preliminary discussions were also initiated with the RATOM Project (http://ratom.web.unc.edu) but divergent timelines did not allow in depth analysis and consideration of that project's outputs by this task force.

most members of the larger task force felt adrift when presented with the models and usure how to use them.

5. Recommendations and Next Steps

The work completed by the Email Archiving Task Force successfully identified requirements for tools used to archive email; the most prevalent tools being used for this work within the archives community; and finally suggested workflows for using these tools during the pre-acquisition, acquisition, accessioning, and access stage of email archiving work. As mentioned in the previous section, the next logical step is for YUL/M units to test the identified tools and workflows using real-world examples from within YUL/M's collections.

In order to move these findings from theory to practice within YUL/M, the task force would suggest the following:

- Members of this task force agree to informally continue testing of these tools and workflows using materials within collections they have access to. The task force will then schedule quarterly check-ins to review this testing and discuss any new thinking. Members of the task force will report any notable findings to BDAWG.
- A brief announcement to the YUL/M archives community about this work
- Expedited completion of an ongoing project to create a centrally hosted and supported suite of software tools for staff working with born digital collections, available to multiple users from multiple units across YUL/M/M. BDAWG members David Cirella and Jonathan Manton are currently working with Library IT to move this project towards a testing phase. The proposed solution would incorporate hosting required tools in an instance of Microsoft Azure DevTest Labs. The principal benefit is that this solution would provide robust access to these tools centrally without the need for each YUL/M unit to maintain licences to required software along with the necessary hardware to run them on locally. This would remove a significant barrier for units, notably those with fewer staff resources, looking to work more actively with their born digital collections, including email. More details on this proposed solution can be found in Appendix E. It should be noted that BDAWG will need to fund the maintenance of this solution for the near future, so should budget for this accordingly.
- The formation of a group of expert users from across YUL/M that could provide ad-hoc advice and consultancy for units across YUL/M for issues related to born digital collections. This would incorporate not only email content, but any born digital collection materials acquired by YUL/M repositories. This mentorship group would be a subgroup of BDAWG, though would not required that members of the group also be current members of BDAWG.
- Further follow up discussions and information sharing with the RATOM project team.

- Investigation, planning and proposing a project to create the custom programming outlined in some of the proposed workflows.
- Explore training opportunities when appropriate, including possibly hosting an instance of SAA's forthcoming course on EPADD at YUL.

Appendices

Appendix A - Core requirements with matching tools
Appendix B - User Stories
Appendix C - Software overview
Appendix D - Process Workflows
Appendix E - Multi user / multi use machine for born digital collection work.