Preservation Workflows and Access Strategies for Donor Correspondence and Documentation: A Final Report

Anne Kofmehl
Capstone, Harry Ransom Center
July 27, 2014

Table of Contents

I. Introduction

In the summer of 2014, I spent two months developing preservation workflows and access strategies for donor correspondence, in both paper and electronic form, and lab notes created within the context of the born digital manuscript collections at the Harry Ransom Center. I was supervised by Lisa Snider, the electronic records archivist for the duration of my project, which took place from June 1st to July 29th. My task for the two months was to review the donor documentation and design workflows from acquisition to access for each component (electronic, paper, and lab notes). In addition to answering the how, what, when, and where questions of preservation, I also considered the implications of access. Who should have access and why? What is the value of this material to an archivist? to a researcher? In order to answer these questions, I had to consider important issues such as confidentiality, privacy, and researcher value.

This report is an answer to those questions and a culmination of my experiences. It contains the following sections: a summary of my research and interviews, a detailed report of my process testing various email preservation programs and creating preservation workflows, recommendations for preservation and access to this material, and concluding thoughts on future implications and directions. In addition, the appendix includes copies of the workflows in diagram and document form, weekly progress reports, sketches and notes, and best practices.

II. Getting Started: Research & Interviews

My first task was to review the donor files from the previous electronic records archivist, Gabby Redwine. Reading through the correspondence helped to clarify the relationship between the archivist and the donor. It also shed light on the amount of work born digital archivists put into gathering archival material from their donors (a lot of trial and error!). After reading through the emails, I began to understand the importance of this material in understanding the archival process and the archivist's job. Not to mention, a new understanding of how having access to this correspondence could mean a great deal to future archivists in this position and others within the department.

In order to familiarize myself with the current digital preservation landscape as it pertains to email and donor documentation preservation, I read several studies on email preservation projects from a variety of institutions. Despite the fact that email has been around for decades and continues to become an increasingly important means of communication in our daily and professional lives, the practice of preserving it is still relatively new and under published.

It is a topic that involves multiple disciplines, from museum curators to records managers, archivists to business professionals. As with many issues of digital preservation, finding a common ground and a universal accepted standard is nearly impossible. While many of the projects I came across have focused on XML-oriented solutions as a means to preserving emails (CERP, Xena etc.). Christopher Prom, who wrote an extensive report on the subject,

points out that while XML provides format neutrality, it does little in terms of providing access (Prom, 23). In addition to providing insight into the various preservation formats tested and used, Christopher Prom's report, "Preserving Email", provided the best insight into the various tools available and a good amount of background history on the subject. I found it most useful as an overall introduction to the field. Maureen Pennock's "Curating Emails:", also provided a useful amount of background knowledge about file formats, email authenticity, and project planning.

Other studies I read, like the Collaborative Electronic Records Project (CERP), a joint effort between the Rockefeller Archive Center and the Smithsonian Institution Archives provided insight into the process of planning a project of this scale, from workflow creation to inception. The University of Manchester Library published a similar report in May 2012 on their work with the Carcanet Press Email Preservation Project. Both reports helped shape the process of this project and provided a useful template for writing this report. A full list of sources consulted throughout this project is available in the bibliography.

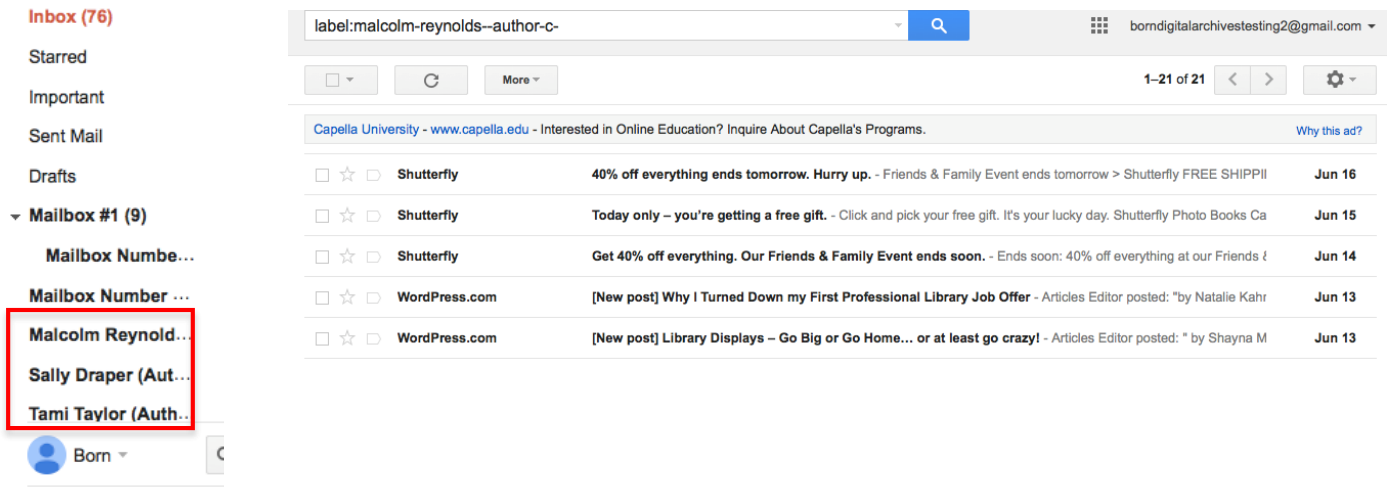
In addition to reading these various reports, I also interviewed the electronic records archivist, in order to understand her own philosophy on email preservation. We discussed what file formats she prefers for access and preservation, what programs she's used in the past, as well as her email habits. This last piece was important in understanding what type of organizational structure was used in relation to the donor correspondence (folders, labels etc.). After researching and interviews, I moved onto the next step of this project: testing and workflow creation. The next section details that process.

III. Process: Preservation Testing and Workflow Creation

a. Email Correspondence

After reviewing past email preservation projects from a variety of institutions, I conducted my own investigation into the various tools and methods they utilized. The following is a summary of my experience broken down by tool, method, and/or program.

Before my testing began, I first had to set up a test email account. I used the Gmail account borndigitalarchivestesting2@gmail.com that I was given by the Electronic Records Archivist. In order to make the test a "real-world" scenario, I created three author labels and filled each folder with 20 emails.



Gmail's Archive Data Tool
http://gmailblog.blogspot.com/2013/12/download-copy-of-your-gmail-and-google.html

Gmail released this tool in December 2013. It allows Gmail users to backup data from their Gmail and Google calendar in a few easy steps. In this scenario, I chose to backup only the author folders. The emails were made available in the mbox format as a compressed folder, which I could download directly from Gmail. This tool was by the far the easiest for completing the simple task of getting the email securely onto your computer. The one drawback to this method is that it does not allow for filtering beyond selecting specific folders. Thus, this method was not recommended due to the fact that filtering is not an option.

Got Your Back
https://github.com/jay0lee/got-your-back/releases

This program, similar to Gmail's Archive tool, allows the user to backup data from their Gmail account using the command line. The program and subsequent code is made available through GitHub. This tool is more hands on and required careful attention to the instructions. While it may require more work, it does have the added benefit of being more flexible and customizable. Unlike Gmail's tool, with Got Your Back you can filter by more than just folder level. Using Gmail's search terms, I created a command that filtered the messages within each folder by date:

> search "label:sally-draper--author-a- AND newer_than:30d"

The messages were saved as individual eml files in a folder on the desktop. The file naming and structure for these emails were hard to decipher. The folder is broken up into a series of numbers, and the messages themselves are given a 5-digit number. If I were to explore this tool further I would probably investigate the file naming and find a new, more suitable, naming convention.

There are other tools available for getting Gmail onto your computer (many of them proprietary in nature) but these two tools were free and easy to use and offered the features necessary to the task. After successfully downloading Gmail onto my computer, I tried a few different tools for converting the emails into a variety of formats for preservation and access.

Emailchemy
http://www.weirdkid.com/products/emailchemy/

The first tool tested was Emailchemy, a product of Weird Kid Software. This program allows for the conversation of email from various proprietary formats into more standard formats like mbox and eml. I had already downloaded emails from Gmail into both of these formats, so this tool was not useful in this case. I converted some of the eml messages that I had downloaded using Got Your Back into mbox and vice versa, but that was the extent of testing. This tool would probably be better suited for Microsoft Outlook or Mozilla Thunderbird clients.

Xena
http://xena.sourceforge.net/

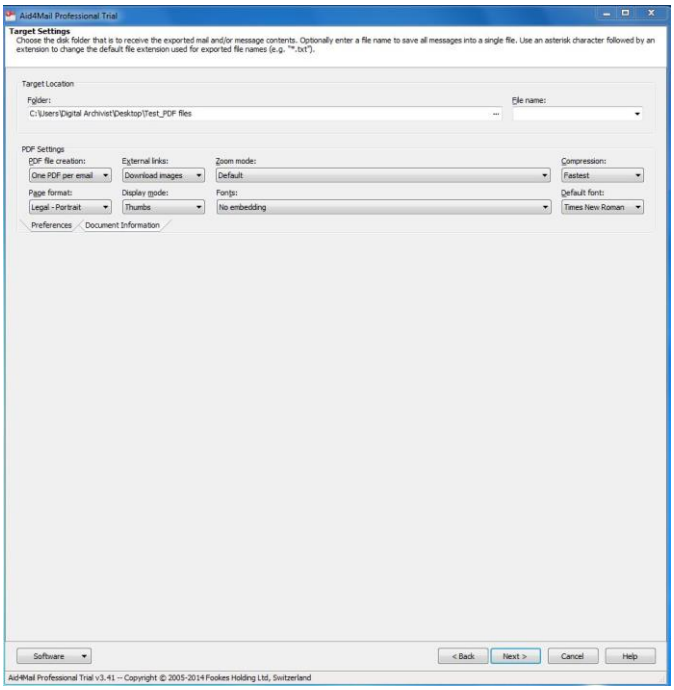
Xena is a free and open source software program created by the National Archives of Australia to aid in the digital preservation of a variety of digital records. It detects a variety of file formats and converts them into open preservation formats, primarily xml. In my research, I came across a couple of groups that have used this program to convert email formats like mbox and eml into xml. Unfortunately, I was unable to get this program to work properly. The version available was outdated and one of the necessary plug-ins failed to work. Therefore, without further testing, this tool would not be recommended.

Aid4Mail
http://www.aid4mail.com/

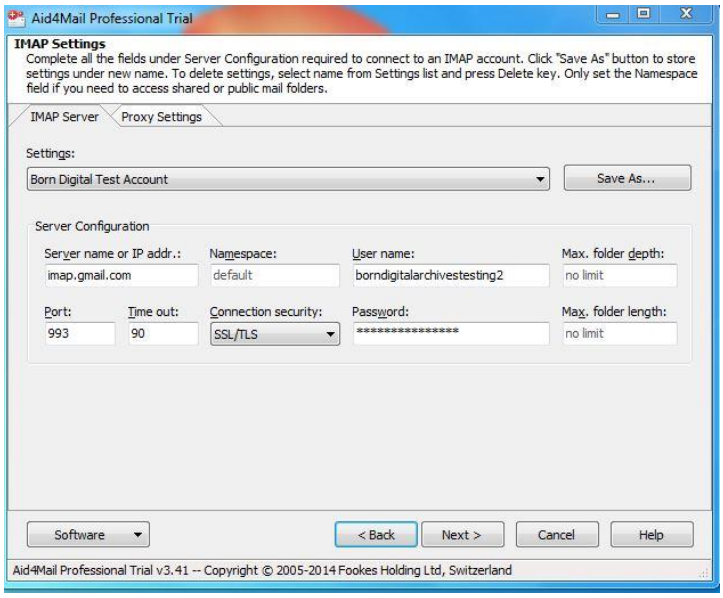
Aid4mail is a proprietary software program that includes the following features: email extraction, conversion, analysis and archiving. It is by far the most robust program of the group tested. It covers a wide range of formats for both access (i.e. pdf) and preservation (i.e. xml). A full list of formats covered can be found on their website. It requires Windows, though it can also run using Linux in conjunction with Wine (an open source software program for running Windows on other operating systems).

Because of its diverse set of features, several scenarios were tested.

Scenario 1: Using mbox files from a previous test (the Gmail archive tool), the files were converted into eml and then converted into pdfs. There are a few different options when converting to a pdf. An entire folder of emails can be converted into one pdf or each individual email can be converted into its own pdf.



Scenario 2: The IMAP option was used to connect directly to the Gmail account to extract the emails. The inbox was filtered by folder first (only the author folders were selected), and then by date (no messages older than 5/31/14). The emails were then converted into the mbox format and into one pdf (per author folder).



Scenario 3: Converted eml and mbox files into the xml format. There are a couple different options for xml. One type includes all the message data, headers and body. The other allows one to filter out the body of the messages and only save the header information as an xml file.

The features, flexibility, and ease of use make Aid4mail the recommended program for the task at hand. The advantage of having a comprehensive program that combines email extraction with conversion outweighs the cost. While other programs tested, like Got Your Back and Emailchemy were efficient, for the purposes of this project Aid4mail would be the best program to use.

After testing, a few different workflows were sketched (see Appendix C). After consulting with the Electronic Records Archivist, the workflow using Aid4mail was chosen and further drafts were made.

The resultant workflow covers the following steps: creation, capture, conversion, transfer, storage, and access. The workflow was designed with the following framework in mind:

On the 1st of each month, the archivist will use Aid4mail to extract the appropriate correspondent folders from their Gmail inbox. The folders will then be converted into mbox, pdf, and xml formats (metadata only) respectively. The files will be saved in a corresponding folder on the desktop. The mbox and pdf files will be transferred to Archivematica for normalization, quarantine, and AIP and DIP creation. The AIPs will be stored in a 3 tier storage system (yet to be determined) and DIPs will be stored in either

ContentDM or AtoM. Please refer to the next section on recommendations to learn more about access restrictions.

For complete written and diagram forms of the workflow please refer to Appendix A & B.

b. Paper Correspondence

In addition to the electronic correspondence, there are approximately 13 folders of paper correspondence from the previous Electronic Records Archivist. In order to make this correspondence easily accessible and searchable, the folders will be scanned using Adobe Acrobat Professional, OCRed, and saved as pdf files in a corresponding folder on the archivist's desktop. The files will then be transferred to Archivematica for normalization, quarantine, and AIP and DIP creation. The AIPs will be stored in a 3 tier storage system (yet to be determined) and the DIPS will be stored in either ContentDM or AtoM. Please refer to the next section on recommendations to learn more about the access restrictions.

For complete written and diagram forms of this workflow please refer to Appendix A & B.

c. Lab Notes

In addition to the electronic and paper correspondence, the archivist's lab notes will also be preserved. Lab notes document the archivist's research and thought process recorded within the context of retrieving archival materials from various donors. The current electronic records archivist's records her lab notes via email, therefore in order to preserve this material a similar workflow to the electronic correspondence will be used with one exception. After saving the emails as pdfs, those emails that contain lab notes will be renamed. The letters 'lab' will be added to the end of the file name to differentiate them from the other correspondence in the author's folder. The paper lab notes will follow the paper correspondence workflow, with the same letters 'lab' being appended to those files that contain lab notes.

For complete written and diagram forms of this workflow please refer to Appendix A & B.

IV. Recommendations

a. Preservation

The following are recommendations for the preservation of the *electronic donor correspondence* from creation to storage.

Creation

- Every correspondent should receive his or her own label in the archivist's Gmail.
- All messages between the archivist and correspondent should be marked with the appropriate label upon receipt.
- All lab notes written in Gmail should also be marked with the related correspondent.
- In the event that the archivist reaches the allotted inbox memory, emails may be discarded if necessary.

Capture

- On the first of each month the archivist will connect to their inbox via Aid4Mail and download donor messages from the previous month.
- If there are no new messages to download, the archivist can skip that month.
- Depending on the frequency or infrequency of donor correspondence, the archivist can adjust the timeline to suit the workload demand.

Conversion

- After downloading the messages using Aid4Mail, the archivist should use the following naming convention for each file saved to the desktop:
  lastnameofcorrespondent_dateoftransfer.file extension
- Each message will be saved as an mbox, pdf, and xml (metadata only) file.

Transfer

- The archivist should create a folder on his or her desktop for each donor. In order to reduce clutter, the folders can be combined into one folder labeled Donor Correspondence_Year.
- The archivist will save files created using Aid4mail to corresponding desktop folder.
- Copies of the mbox and pdf files will be transferred to Archivematica for AIP and DIP creation**.**
- In the event that Archivematica is not used, an equivalent substitute is acceptable.

Storage

- For long-term storage, preservation copies (AIP) created in Archivematica will be stored in a 3 tier storage system that is yet to be determined.
- For mid-term storage and user access, access copies (DIP) created in Archivematica will be stored in ContentDM, AtoM, or an equivalent.

The following are recommendations for the preservation of the *paper correspondence* from creation to storage.

Creation

- The folders should be labeled clearly with the donor's name and date range of correspondence. The folders should be kept securely in the Electronic Archivist's office, preferably in an archival safe storage box.

Capture

- The contents of each folder should be scanned and OCRed using Adobe Acrobat Professional and saved as pdfs to the desktop in a similar fashion as the electronic correspondence.

Conversion

- When saving the files the following naming convention should be used for the files: *lastnameofcorrespondent_dateofemail_page#(if applicable).pdf*

Transfer

- The archivist should create a folder on his or her desktop for each donor. In order to reduce clutter, the folders can be combined into one folder labeled Donor Correspondence_Year.
- The archivist will save the digitized emails to corresponding desktop folder
- Copies of the pdf files will be transferred to Archivematica for AIP and DIP creation.
- In the event that Archivematica is not used, an equivalent substitute is acceptable.

Storage – Mid and long term storage

- For long-term storage, preservation copies (AIP) created in Archivematica will be stored in a 3 tier storage system that is yet to be determined.
- For mid-term storage and user access, access copies (DIP) created in Archivematica will be stored in ContentDM, AtoM, or an equivalent.

Other considerations

- The timeline for completing this portion of the project is flexible. One method would be to scan and OCR all the folders at once. The other method would be to only scan those folders that are of immediate relevance (i.e. donors with whom correspondence is ongoing). The archivist should use their discretion in deciding the timeline.
- The digitization of the paper correspondence can be carried out by another staff member, if necessary.

The following are the recommendations for the preservation of the *lab notes*

For the lab notes, the recommendations are the same as above, with one exception.

- PDF files containing lab notes should be renamed to reflect this difference. The letters 'lab' will be added to the file so it will read as such: *lastnameofcorrespondent_dateoftransfer_lab.pdf*

These recommendations are subject to change based on the experience and resources of the archivist. In six months the workflow should be reviewed and any inefficiencies should be addressed and changed as necessary. At that six month mark, files saved on the desktop and in storage should be checked against the originals to determine any loss.

b. Access

Access to donor correspondence and lab notes will take place in two stages. The first stage restricts access to internal stakeholders only. The second stage would open it up to a wider audience. The second stage can be considered optional. It is still unclear whether or not material of this nature will be of any interest to researchers. Granting access to donor correspondence and archivist's lab notes is not a common practice among archives. One of

the larger concerns is of course privacy and confidentiality. Correspondence among archivists and their donors often consists of sensitive and personal information. Prior to granting access to this material it would be prudent to consult the donor, as one can imagine they would not expect this material to be made available publicly.

On the other hand there is the potential that material of this nature could be of great interest to researchers, specifically I believe other archivists from like-minded institutions might be interested in understanding the relationship between donors and archivists and how business is conducted. We could make significant advances by sharing methods, triumphs, and failed experiments with other archivists in this new and evolving field of born digital archiving.

The key here is to understand the researcher and their interests. It might be worth looking at this experience as an opportunity to expand that understanding. If access is granted, steps should be taken to survey users in order to understand their motivations behind studying this material. What insight, if any, do they gain from this material that they can't get elsewhere?

1st stage – Internal Access Only

Due to concerns of privacy and confidentiality, the correspondence and lab notes will be restricted to the following individuals:

Donor
Electronic Records Archivist, Lisa Snider
Electronic Records Archivist's Supervisor, Stephen Mielke
Associate Director for Acquisitions and Administration, Megan Barnard

Other staff members of the Harry Ransom Center may also gain access on a case-by-case basis. Request for access will need to be cleared by both the archivist and his or her supervisor.

2nd Stage – Researcher Access

In the event that the archivist, with the permission of both the donor and his or supervisor, decides to grant access to researchers the following steps should be taken:

- An inventory of emails containing basic information such as subject title and date will be made available to the researcher.
- The researcher can request specific emails.
- Emails should be vetted for sensitive information and information deemed as such will be redacted. The donor should be consulted for final approval.
- A secure environment should be set up in which the researcher can view the material, in the case of the Harry Ransom Center this will most likely be a computer station set up in the reading room.
- Usage statistics should be recorded to determine what types of emails are frequently requested.

V. Best Practices

The following are guidelines for setting up folder structure, file naming, file formats, and digitization of paper files.

Folder Structure

- Folders should be named with major functions (i.e. Donor Correspondence 2014)
- Organize sub-folders by Month / Year
- Use last names, first initials for correspondents (i.e. Coetzee, J.)
- Keep names simple, consistent, and self-explanatory

File Naming

- Use correspondent last name and first initial if necessary when naming files
- Avoid illegal characters: > < " / \ | ? * : ^ $
- Avoid spaces, use underscores instead (i.e. Coetzee_20140731.mbox)
- Capitalize the first letter in each word (i.e. Coetzee_20140731_Lab.pdf)
- Use YYYYMMDD or YYYY-MM-DD for dates (20140731 or 2014-07-31)
- Examples:
    - Coetzee_2014-07-31.mbox
    - McEwan_2014-06-30.xml
    - Coetzee_20140715_Page1.pdf
    - Coetzee_2014-07-31_Lab.pdf

File Formats

- Email Correspondence
    - MBOX (Preservation Copy)
    - PDF and PDF/A (Access Copy)
    - XML (Metadata)
- Paper Correspondence
    - PDF and PDF/A (Preservation and Access Copy)
- Lab Notes (Electronic)
    - MBOX (Preservation Copy)
    - PDF and PDF/A (Access Copy)
    - XML (Metadata)
- Lab Notes (Paper)
    - PDF and PDF/A (Preservation and Access Copy)

Digitization Guidelines

Adobe Acrobat Professional has a built in scanning tool for scanning print materials as PDFs. Use the following steps for creating PDF files using Adobe Acrobat Professional[1]

---

[1] These steps are adapted from University of Michigan's Best Practices for producing high quality PDFs, accessed July 27, 2014 from http://deepblue.lib.umich.edu/bitstream/handle/2027.42/58005/PDF-Best_Practice_v3_CC-BY.pdf?sequence=45

- From the **File** menu, select **Create > PDF from scanner**. Select your scanner device, choose **Front Sides** or **Both Sides** as appropriate, select **Recognize Text Using OCR**, and select **Add Tags To Document**.
- Click **Image Settings,** and a new box will appear. Use the following settings:
    - Color / Grayscale: JPEG
    - Monochrome: CCITT Group 4
    - High Quality
    - Deskew: Automatic
    - Background Removal: Low
    - Edge Shadow Removal: Cautious
    - Despeckle: Low
    - Descreen: Automatic
    - Halo Removal: On
- Click **Scan**

VI. Conclusion

The recommendations and guidelines for preservation and access of donor correspondence and lab notes in this report are subject to change. After six months, the archivist should formally review the process in order to approve on the existing workflows. In addition, an informal survey should be conducted of users in order to gain insight into how this material is being used. Prior to moving forward into the second stage of access, the archivist should test possible avenues of access (reading room, online etc.) to find the best fit for this material. Consideration should be given to issues of confidentiality and privacy, and sensitive information will have to be redacted.

In the future, steps should be taken to study the implications of extending this practice into other departments at the Harry Ransom Center. Other manuscript collection archivists and the director of acquisitions could benefit from a similar process for capturing and preserving correspondence with donors. . In order to create more awareness within the archival community surrounding these issues of internal documentation, efforts should be made to share results of this project and its future stages with the wider community through publications and conferences. The hope is that by sharing this information, as a community, we can take steps to insuring that future archivists and researchers have access to this important material. By preserving this documentation, we remove the fourth wall, and give them the opportunity to reflect upon the archival process and learn from the trials and triumphs the past.

Appendix A: Workflows

BORN DIGITAL MANUSCRIPT COLLECTION

Donor Correspondence Preservation Workflow (digital) *created by Anne Kofmehl, July 2014*

CREATION

Gmail Inbox

Step 1: Create a label in your Gmail inbox for each author/correspondent.
Step 2: As emails are received, save them in the appropriate labeled folders.

Desktop

Step 3: Create folders for each author/correspondent.

CAPTURE

Step 4: Open Aid4Mail
Step 5: Connect to your Gmail account via IMAP connection
Step 6: Select ONLY the author/correspondent folders you wish to transfer
Step 7: Filter by date*

CONVERSION

Step 8: Convert each author/correspondent folder into an MBOX file, save as *author.lastname_dateoftransfer.mbox*
Step 9: Convert each MBOX file into a PDF (save as individual email per pdf), save folder containing pdfs as *author.lastname_dateoftransfer PDFs*
Step 10: Create a Metadata XML file (saves only the header information) for each author/correspondent MBOX file, save as *author.lastname_dateoftransfer.xml*

TRANSFER

Step 11: Transfer MBOX, PDF, and XML files to corresponding folders on desktop (*for internal access*)
Step 12: Transfer MBOX and PDF files to Archivematica for AIP and DIP creation

STORAGE

Step 13: AIPs (MBOX and PDF) are stored in *HRC New Storage*
Step 14: DIPs (MBOX and PDF) are stored in *CONTENTdm* or *AtoM*

FUTURE CONSIDERATIONS

Access: Initial access will be internal (primarily Lisa, and other interested parties)
In the future, if researchers are allowed access it will be in the reading room only, with web access to come later.

REPEAT STEPS 4-13 on the 1st of each month**

---

* If it is July 1st, filter out anything older than 6/1, so on and so forth.

** This is just a suggestion to start, if the project is completed within a couple of weeks, feel free to adjust the schedule as needed. The idea here is to establish some sort of routine, and a month feels the most appropriate; but in practice that might change.

BORN DIGITAL MANUSCRIPT COLLECTION

Donor Correspondence Preservation Workflow (paper) *created by Anne Kofmehl, July 2014*

CREATION

Desktop

Step 1: Create folders for each author/correspondent.

CAPTURE

Step 2: Open Adobe Acrobat Professional and Scanner Program

Step 3: In Adobe Acrobat Professional, from File menu, select **Create > PDF** from Scanner, select **Recognize Text Using OCR**, and select **Add Tags to Document**

Step 4: Scan emails in each folder as PDFs, maintaining the thread order.

Step 5: Check OCR

CONVERSION

Step 6: Convert each page into a pdf, save as *authors.lastname_dateofemail_page #(if applicable).pdf*

TRANSFER

Step 7: Transfer PDF to corresponding folders on desktop (*for internal access?*)

Step 8: Transfer PDF files to archivematica for AIP and DIP creation

STORAGE

Step 9: AIPs (PDF/A) are stored in *HRC New Storage*

Step 10: DIPs (PDF/A) are stored in *CONTENTdm* or *AtoM*

FUTURE CONSIDERATIONS

Access: Initial access will be internal (primarily Lisa, and other interested parties)
In the future, if researchers are allowed access it will be in the reading room only, with web access to come later.

Special Note: This is the ideal workflow, alternatives would be to scan on a need-to-know-basis (i.e. authors that you are currently working with like Coetzee), OR you could not scan at all, and keep them as paper files. I think any option would be acceptable.

BORN DIGITAL MANUSCRIPT COLLECTION

Lab Notes Preservation Workflow (digital) *created by Anne Kofmehl, July 2014*

CREATION

Gmail Inbox

Step 1: Create a label in your Gmail inbox for each author/correspondent.

Step 2: Copy lab notes into an email, save them in the corresponding author/correspondent folder.

Desktop

Step 3: Use the same author/correspondent folders created for the digital correspondence.

CAPTURE

Step 4: Open Aid4Mail

Step 5: Connect to your Gmail account via IMAP connection

Step 6: Select ONLY the author/correspondent folders you wish to transfer

Step 7: Filter by date*

CONVERSION

Step 8: Convert each author/correspondent folder into an MBOX file, save as *author.lastname_dateoftransfer.mbox*

Step 9: Convert each MBOX file into a PDF (save as individual email per pdf), save folder containing pdfs as *author.lastname_dateoftransfer PDFs*

Step 10: Rename the lab note files by adding the letters LAB to the end. It will look like this: *author.lastname_dateoftransfer_lab.pdf*

TRANSFER

Step 11: Follow the steps under digital donor correspondence.

STORAGE

Step 12: Follow the steps under digital donor correspondence.

FUTURE CONSIDERATIONS

Access: Initial access will be internal (primarily Lisa, and other interested parties)

In the future, if researchers are allowed access it will be in the reading room only, with web access to come later.

REPEAT STEPS 4-13 on the 1st of each month**

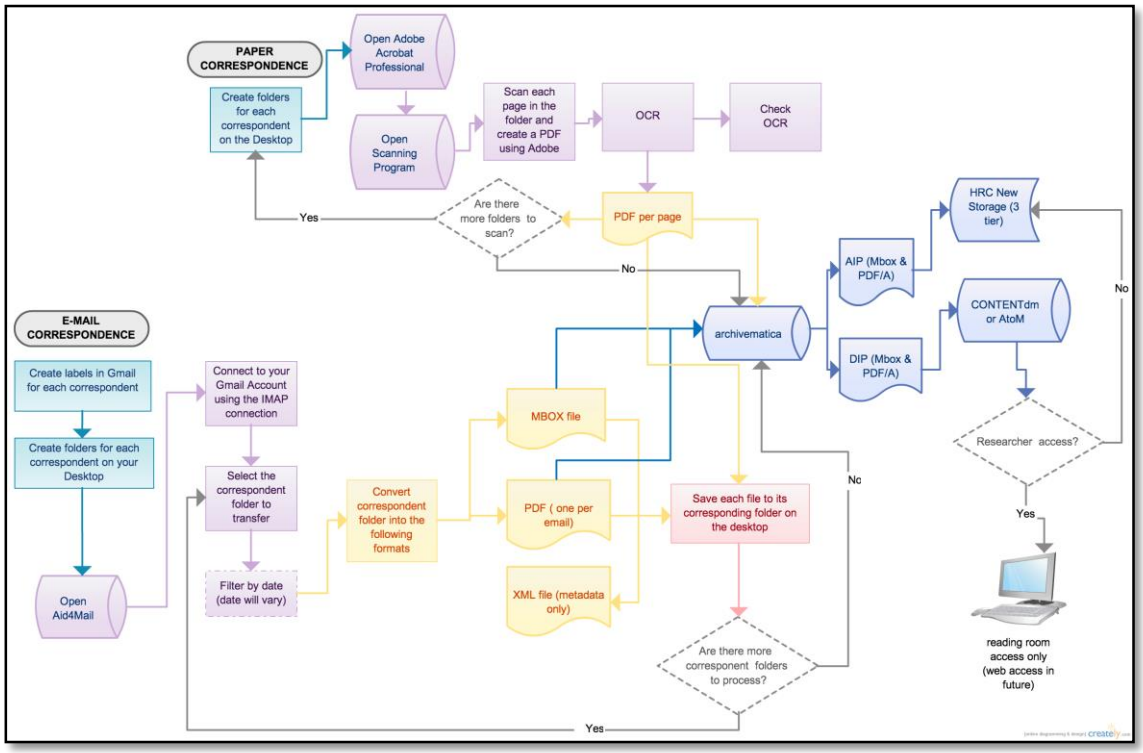
---

* If it is July 1st, filter out anything older than 6/1, so on and so forth.

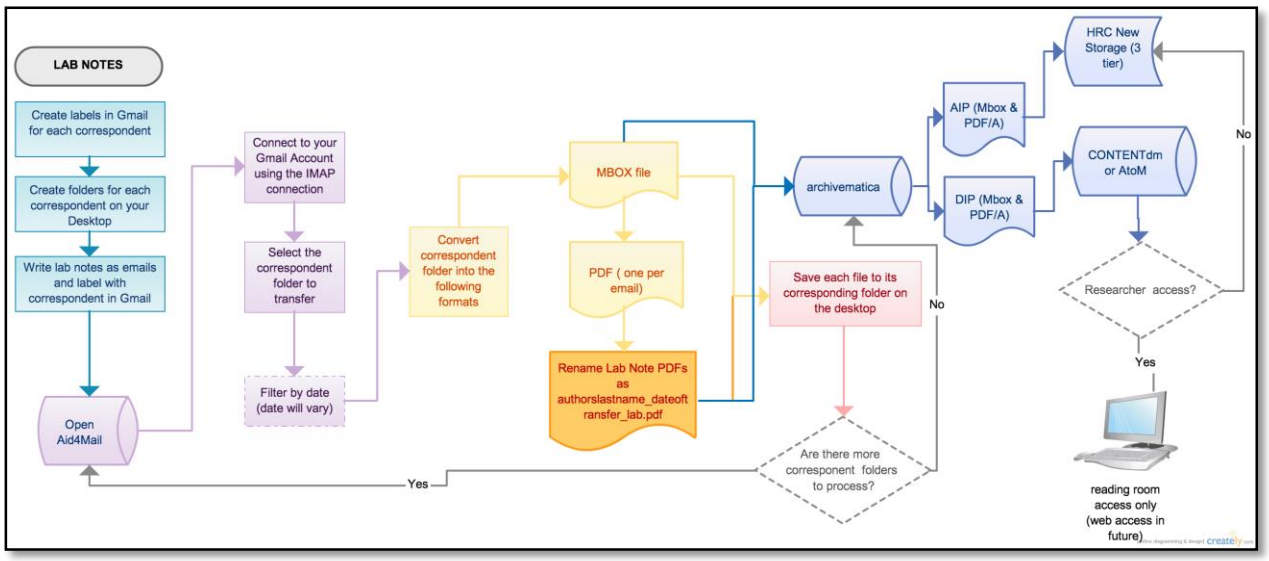
** This is just a suggestion to start, if the project is completed within a couple of weeks, feel free to adjust the schedule as needed. The idea here is to establish some sort of routine, and a month feels the most appropriate; but in practice that might change.

Appendix B: Workflow Diagrams

Workflow 1: Email & Paper Correspondence



Workflow 2: Lab Notes



Appendix C: Workflow Drafts

WORKFLOW SKETCH # 1 -- Aid4Mail
Create "Author" Labels in Gmail
(In my example: Sally Draper_Author A
Tami Taylor_Author B
Malcolm Reynolds_Author C)
On the desktop create folders for each author (save all copies: MBOX, PDF, etc. to the corresponding author folder)
Open Aid4Mail
Connect to gmail via IMAP connection
Select ONLY the author folders
Filter by date (Once a month? Or more frequently, weekly?)
Convert Folders to MBOX - Save files
Convert MBOX files to PDF (one pdf per folder)- Save files OR Convert to EML and then to PDF (individual emails) and save files
Zip each author file?
-draper_6-30-14.MBOX
-draper_6-30-14.PDF or draper_[date of individual email].pdf
-eml messages?
--------> then to Archivematica??
questions to consider: Naming conventions? i.e.: lastnameauthor_date of most recent download.mbox OR _date range? Metadata-- where and what?

WORKFLOW SKETCH #2 -- Gmail Data Archive, Aid4Mail
Open Gmail's Archive Tool
Select only the Author Folders
Download and unzip MBOX file folder and save it to the desktop
On the desktop create folders for each author (save all copies: MBOX, PDF, etc. to the corresponding author folder)
Open Aid4Mail
Convert Folders to MBOX - Save files
Convert MBOX files to PDF (one pdf per folder)- Save files OR Convert to EML and then to PDF (individual emails) and save files
Zip each author file?
-draper_6-30-14.MBOX
-draper_6-30-14.PDF or -draper_[date of individual email].pdf
-eml messages?
--------> then to Archivematica??

WORKFLOW SKETCH #3 -- Aid4Mail (include an XML step?)
Create "Author" Labels in Gmail
Open Aid4Mail
Connect to gmail via IMAP connection
Select ONLY the author folders
Filter by date (Once a month?)

[preservation copy] Convert Folders to MBOX - Save files
[access copy] Convert MBOX files to PDF (one pdf per folder)- Save files OR Convert to EML and then to PDF (individual emails) and save files
[XML version] Convert MBOX to XML or Convert EML to XML
Zip author file
-draper_6.30.14.mbox
-draper_6.30.14.pdf
-draper_6.30.14.xml
-eml messages?
--------> then to Archivematica??

Paper Correspondence Workflow Sketch # 1
Scan emails
OCR
Save as PDFs in Correspondent Folders on Desktop
---feed into the other workflow
Transfer to Archivematica -- AIP and DIP created
Store in HRC New Storage / ContentDM

Appendix D: Progress Reports

6.9.14

HOURS WORKED 6/2-6/6: 8 hours

Progress Report

This week I spent time reading through the files Gabby Redwine kept for the various donors she worked with. It was interesting to see the variety of dialogue that occurred between her and the authors who were donating their materials, as well as other staff members at the HRC. Some files were rather thin, while others were more robust (i.e. Coetzee). Not knowing Gabby or her methods / system for working with donors, these folders (albeit at times incomplete), offered an important insight into that process. I'll be interested to see how your correspondence and notes will compare with hers.

Questions that have popped up so far, that I'll be interested to explore further are as follows:

1) What is the value of this material? To an archivist? And perhaps to a researcher? (this is a tricky one)
2) How should it be kept? Print outs? Electronically? (I found it sort of tedious to leaf through so many sheets of paper, especially when it comes to long chains of email, it can be a bit redundant / hard to follow)
3) In the larger scheme of things: how have other institutions dealt with their donor correspondence? Is there a precedent I can look toward, in terms of workflows / methods etc.

I think the next step will be to look through what you have collected and then to do some research specific to that third question above (what have others done) and after that compile some questions and maybe interview some people (i.e. the email you just sent me).

Let me know if I'm on (or off) the right track!

6.15.14

HOURS: 14.5 (3.5 hours spent at home)

This week I've been reading through several long reports on email preservation, including Prom's *Preserving Email* and Pennock's *Curating E-mails*. I'm currently reading the CERP report, but as of now have yet to finish it. Overall, I've found the documentation very helpful at painting a broader picture of email preservation and its current stance in the archivist's world. The readings have also raised a lot of helpful questions/thoughts that I think will be useful in moving this project forward. Some of those questions/thoughts are as follows:

1) Does HRC have an email management policy for its employees? Are there any policies in place for migrating staff-to-staff and staff-to-donor correspondence to short-term or even long-term storage space? (Perhaps this is where talking to Megan would be useful?)

2) I was interested in the various email preservation tools Prom mentioned in his report. Aid4Mail, MailArchiva, etc. (p.26-27) I'm wondering if it would perhaps be helpful to test some of these programs using the dummy email account you set up? (We can talk about this more in person I think)

3) XML vs. PDF for preservation? Prom drove it home most clearly, but others have alluded it to as well, as XML being the best best in terms of preservation. Your thoughts? There are also mentions of various formats like MBOX, EML (I think it was Prom who mentioned these formats are best for access)

Moving forward, I think I have a little more reading to do (maybe you have more suggestions?), but I'd also like to get started testing some of these programs out, maybe sketching a few potential workflows? I also think it would be helpful to conduct some interviews (with Megan, Gloria etc.), but I'm not sure when would be the best time to do so (I'm not quite clear on what kind of information I could be gleaning from them). Since this project (at least in its initial stages) is focused on your email correspondence, I thought it might be helpful to conduct a survey/interview of sorts to gain a better understanding of your email habits (i.e. in the spirit of CERP's survey they conducted with the email users they were testing).

So bearing all of that in mind, I was thinking that a good goal for the end of this month would be to have a rough draft of a preservation workflow (I think this was the first deliverable we talked about). Does that sound reasonable to you?

6.30.14

HOURS WORKED 6.16-30: 21 hours

After our conversation about the donor correspondence and a little more research on my part, I spent this past week testing out various methods / programs for preserving emails. I'll go through them one by one and share my thoughts on each, and offer some conclusions at the end.

Before I got started with any of the programs, I went ahead and organized a portion of the emails within the test account into author folders (and named them after favorite tv characters :)). Each folder had 20 emails in it to start, I added others as I went through the different program tests.
These are the names of the folders:

Sally Draper (Author A)
Tami Taylor (Author B)
Malcolm Reynolds (Author C)

The first thing I did was use **Gmail's Archive Data Tool** to extract the emails from those specific folders. This was *really* simple and because there weren't many emails it was a quick download time. The total file ended up being 386KB in size.

Next, I used **GYB**, the command line tool, to backup the gmail account. This was more complex and a little more of a challenge, but I did like that it offered a bit more flexibility than Gmail's archive tool. After getting the account set up initially and doing an overall backup, I used the --search function to limit to just the author folders, and then after that I made the search a little more complex to limit to a specific date range (trying to keep in mind that there may be the need to limit the back-up in such a fashion).
These were the commands I used:

--search "label:sally-draper--author-a-"
--search "label:sally-draper--author-a- AND newer_than:30d"

I looked at this page, to help me figure out how to use the gmail search:
https://support.google.com/mail/answer/7190?hl=en&ref_topic=3394914

After playing around with those two tools for getting gmail onto a computer, I downloaded and installed **Emailchemy**. I didn't spend too much time with this tool because it didn't offer a whole lot in terms of formats. I did convert some EML files (from GYB) into MBOX and vice versa.

I did the same thing with **Xena**, because I had read about it in the two iSchool projects. I couldn't get it to work, so I abandoned it pretty quickly. I had to install an older version because the newer version was not compatible with my Mac.

I went into HRC to use the PC to download/install and run **Aid4Mail.** This was by far my favorite tool of them all. I tried several different scenarios with this one. Before I got started in the program, I used the gmail archive data tool to download mbox files of the three author folders, then I tried the following scenarios:

Scenario 1: I converted the MBOX files to EML files, so each email would have an individual file, and then I converted them to PDFs, with individual pdfs for each email.

Scenario 2: I used the IMAP option to connect directly to the GMail account. I filtered by date (no older than 5/31/14) and then I converted the emails to MBOX. Next I took each mbox file and made it into a pdf (one pdf per folder, not individual email), and also tried converting to XML.

Scenario 3: I continued to play around with the xml function, converting both eml files that I had filtered by date and mbox files into xml.

Overall, I really like the idea of using Aid4Mail. I know it is proprietary software, so I'm not sure how viable it is in real life, but it seems like a great program. A couple of times it

stalled out during the conversion stage (it would say 'not responding', but then come back after a couple of seconds). I think once I had to ctrl+alt+delete and shut it down and start it up again. I didn't try filtering beyond date, but perhaps I could play around with that a bit more, see how powerful the search function--that might be useful. I think the thing that impresses me the most is the amount of formats they support. The fact that you can convert messages to EML, MBOX, PDF, and XML, among others, is really great. I'm going to attach another document that has some rough outlines of a potential workflow(s). You can let me know what you think/if I'm on the right track and offer edits, if needed.

Final Thoughts:

I feel a little unclear as to where to go with this next. I know I need to address Gabby Redwine's correspondence at some point, as well as think more beyond just getting the emails out of your inbox and into a new format/folder (you'll see where I sort of stop with the workflow)--but also think about where they go from there: Conceptually, are they kept with the digital objects in the repository space? or elsewhere? and then of course the access piece needs to be worked through. I think it would be helpful for me (so I know I'm staying on course) to mark out some deadlines for the next couple of weeks (i.e. draft due dates etc.) Also, one last thing, I wrote down some questions for Gloria. I'll bring those with me tomorrow as well. I was thinking I would try to be in touch with her this week. I don't want her to forget about me!

Let me know if you have questions, this was a long report!

7.7.14

HOURS WORKED: 17

This past week I spent some time editing the workflow sketch I submitted to you last week (document attached in email), and researching workflow diagrams (thanks for those papers, they were really helpful!) I used the (free!) desktop version of creately to make my diagram. Here's the link.

As you can see I still have some questions regarding quarantine/virus scanning and the end result (HRC portal? or HRC shared drive?), and where does archivematica fit in (??)

After playing around with archivematica a bit, I changed the workflow slightly to clarify what exactly would be transferred there--in this case just the mbox file. I was also curious how we would handle the DIPs, would it be necessary to create one? If so, what program would you likely be using: CONTENTdm or AtoM?

I'm also not sure how detailed I should get with the Archivematica piece, there is room for a lot of metadata to be added, and I wasn't sure if clarifying what fields are essential etc. is a necessary part of this document.

There's also, of course, the paper correspondence to consider. I haven't had much time to devote to that yet. But I'm thinking when I return I can tackle that portion of the project. I'm imagining two different workflows, or would you rather it be all included into one?

I haven't heard back from Gloria yet, but I'm not worried if she doesn't reply. It would be nice to have her opinion / experience to reflect on, but I think I can come to my own conclusions / recommendations without it.

I'll be out of town this week, but will be back on Tuesday July 15th. Looking ahead I hope to have the paper portion wrapped up by the end of that week (July 18th), so I can spend the remainder of time putting together the final report, creating the poster etc.