Email Archiving: Building Capacity and Community

CNI Spring 2021 Virtual Membership Meeting Project Briefing

Ruby Martinez (rubylm2@illinois.edu)
Matthew Connelly (mjc96@columbia.edu)
Gregory Wiedeman (gwiedeman@albany.edu)
Michelle Gallinger (mgallinger@gallingerconsult.com)

ILLINOIS
Agenda

- Email Archives: Building Capacity and Community
- Creating Email Archives from PDFs: The Covid-19 Corpus
- Mailbag: A Stable Package for Email with Multiple Masters
- CoSA PREPARE: Preparing Archives for Records in Email
Email Archives: Building Capacity and Community

https://emailarchivesgrant.library.illinois.edu

Presenter: Ruby Lorraine Martinez
Email Archives Community Fellow (rubylm2@illinois.edu)

Chris Prom
PI and Project Lead (prom@illinois.edu)
Program Goals

Supporting specific process improvements to increase email archives tool functionality and interoperability.

Develop a community of practice that establishes a baseline of email acquisition, processing, discovery, and delivery services, suitable to a wide range of archival repositories.

Highlighting specific process email collections and demonstrating how other archives can process and provide access to email using community-supported tools.
First Round Overview

Subawards

- Harvard University
- University of Albany, SUNY
- Council of State Archivists, Inc.
- Columbia University
- University of Chicago, Library

Key Details:

- Focus on tool development
- Scalability
- Capacity Building
Second Round Funding

1. Second round applications will be due June 30, 2021
2. Areas of focus:
   a. Community Development
   b. Community of practice with or refinement of existing tools
Creating Email Archives from PDFs: The Covid-19 Test Corpus

Presenter: Matthew Connelly
Principal Investigator (mjc96@columbia.edu)
Project Overview

• Create machine-readable corpus of documents on government responses to COVID-19, gathered through FOIA requests
  • *Facilitate new text and data mining research on the social impact of COVID-19, ensure long-term preservation of records*
• Assemble an interdisciplinary panel of public health experts, historians, journalists, and social scientists to identify the important subjects for follow-on document requests
  • *Community- and researcher-driven collection development*
• Develop open source tool and workflow for making email records released as PDF caches compatible with email archiving ecosystem
  • *Build tools and frameworks reusable by researchers & library and archival professionals -- this is the part that the UI/Mellon Email Archiving Initiative is supporting*
Contribution to email archiving solutions on both ends of the email stewardship lifecycle

- Develop code for tool and associated library that converts PDF email archive records into MBOX format
- Using regex and machine learning-based image processing algorithms, identify email data elements such as date, subject, to/from, body, email threads, and the inclusion of attachment files
Contribution to email archiving solutions on both ends of the email stewardship lifecycle

- Integrate into FOIArchive as new COVID-19 collection for enhanced research access (browse via web UI, download as data via API, discover via Columbia University Libraries’ online catalog); preserve for permanent access in Libraries’ long-term storage program
- Load generated MBOX files into ePADD Appraisal and Discovery module as initial use case for PDF-to-MBOX tool
Collection Scope

The new COVID-19 corpus will incorporate records from the following sources:

- **Documenting COVID** ([https://documentingcovid19.io/](https://documentingcovid19.io/))
  - 200,000+ pages of materials, from FOIA requests filed to state and local government agencies
  - Project led by Derek Kravitz (Columbia Journalism School / Brown Institute for Media Innovation)

- **DocumentCloud/MuckRock** ([https://www.muckrock.com/](https://www.muckrock.com/))
  - Response files from state, local *and federal* FOIA requests filed by DocumentCloud/MuckRock
  - ~100 completed in full or in part, ~500 awaiting response

- **New FOIA requests**, as guided by expert panel
  - Under guidance of the expert panel and based on academic and policy research priorities, file new FOIA requests to gather data and information on areas not covered above
Email records as research resource for government responses to COVID-19

<table>
<thead>
<tr>
<th>The Largest Files So Far</th>
<th>Filecount</th>
<th>Space (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dakota County Nebraska</td>
<td>16</td>
<td>2255</td>
</tr>
<tr>
<td>Seward County Kansas health full</td>
<td>28</td>
<td>824</td>
</tr>
<tr>
<td>Tarrant County, Texas</td>
<td>9</td>
<td>475</td>
</tr>
<tr>
<td>Florida</td>
<td>5</td>
<td>458</td>
</tr>
<tr>
<td>Mohave County Arizona</td>
<td>2</td>
<td>451</td>
</tr>
<tr>
<td>Hotspots Caitlin</td>
<td>5</td>
<td>423</td>
</tr>
<tr>
<td>Coffee County Georgia</td>
<td>6</td>
<td>336</td>
</tr>
<tr>
<td>Pinal County Arizona</td>
<td>370</td>
<td>261</td>
</tr>
<tr>
<td>Rally</td>
<td>6</td>
<td>238</td>
</tr>
<tr>
<td>Madera county</td>
<td>1</td>
<td>226</td>
</tr>
</tbody>
</table>
Gerald Millet

From: Kishia Grayson
Sent: Monday, June 15, 2020 9:09 AM
To: Lisa Brochard; LDOECOVID19Support; Beth Scioneaux
Cc: Sheila Campbell; Andrea Burl; Daisy Grotsma; Kahree Wahid; Melinda George; Valerie Black; Nasha Patel; Shannan Smith (DOE)
Subject: Licensed Centers and Registered Homes as of 6.15.2020 AM

Good Morning-

Below is a recap of closures as of Monday, June 15, 2020 at 9:00 a.m.

Licensed Centers Closed – 524 of 1463; 36% of all licensed centers
Family Child Care and In-Home Providers Closed – 28 of 209; 13% of all Family Child Care Providers and In-Home Providers
552 of 1672 – 33% of all providers are closed

Family Child Care and In-Home Providers
53 new family child care and in-home providers opened since 3/23/2020
- 3 - Acadia Parish – Family Child Care Provider (Opened to provide care for the essential workforce)
- 3 - Ascension Parish – Family Child Care Provider (Opened to provide care for the essential workforce)
- Caddo Parish – Family Child Care Provider (Opened to provide care for the essential workforce)
- Calcasieu Parish – Family Child Care Provider (Opened to provide care for the essential workforce)
- 12 - East Baton Rouge Parish
  - Family Child Care Provider (10 – Opened to provide care for the essential workforce)
  - In-Home Provider (Opened to provide care for the essential workforce)
- 4 - Iberville Parish – Family Child Care Provider (3 – Opened to provide care for the essential workforce)
- 2 - Jefferson Parish
  - Family Child Care Provider (Opened to provide care for the essential workforce)
  - In-Home Provider (Opened to provide care for the essential workforce)
Mailbag
A Stable Package for Email with Multiple Masters

Presenter: Gregory Wiedeman
Principal Investigator (gwiedeman@albany.edu)
Email Preservation

- No single master format for email
  - MBOX/EML don't preserve external content
  - PST are similar and proprietary
  - PDFs do not preserve structure
  - Not all emails are static documents
- Not many open source processing tools usable for archivists
UAlbany,

Happy New Year!

As you know, the 115th Congress just got underway. And, I've recently taken on a new role. I was elected by my colleagues to serve as the Chair of the House Democratic Caucus – the fourth highest ranking position in our leadership.

There are numerous challenges that lie ahead, but I am excited to be a part of the team that will lead the fight to protect the progress we've made over the last eight years and to expand on those successes.

Others seem to have taken notice of my growing role in the House Democratic Caucus. Earlier this week, The Hill named me one of the top rising stars in the House. [Click here to read the full article >>]

Seven rising Democratic stars

Just In...
Friend,

In just 30 hours, we will know the fate of our country.

In just 30 hours, we'll see if our hard work to take back the Senate this year has paid off. But until then there is still more we can do. It's up to us to make sure we don't let the closest races slip through our fingers.

That's why I'm asking you to give your last-minute support to Deborah Ross from North Carolina. She's in the closest race for the Senate right now, and we have to help her to make sure she pulls ahead in time for Election Day. If we win this race, I know we can win the Senate.
Multiple Master Formats

- MBOX/EML preserves structure as data
- PDF preserves document and allows for easy access
- WARC preserves interactivity
A Stable Package

- BagIt
  - Widely used
  - Validates fixity
  - Uses filesystem for structure

- Interoperable
  - ePADD
  - RATOM

Mailbag Specification

![Diagram of a Mailbag containing CSV/SQLite, MBOX/EML email data, PDF files, and Web Archives]
Email Processing for Archivists

- Easy packaging near-to-capture
- Existing tools are challenging for archivists
  - Python email libraries
  - Batch PDF creation from email
  - Web Archives capture
- mailbag Python library
- Command Line utility
- Basic GUI with Gooey
Mailbag: A Stable Package for Email with Multiple Masters
CoSA PREPARE: Preparing Archives for Records in Email

Presenter: Michelle Gallinger
Principal Investigator (mgallinger@gallingerconsult.com)
PREPARE Approach

Needs Assessment
Documents
Testing
Mentoring and Assistance
Deepening Community
Needs Assessment

Survey focusing on the current email archiving capacities of the 56 state, territorial, and District of Columbia archives to determine levels of capability, need, and support for email archiving.

Survey results will identify and prioritize state government records support needed to support:
- Transfer
- Preservation
- Access
Documents

Cite existing preservation literature in specifically tailored guidance documents for the preservation of email in state/territorial archives.

Detail actionable recommendations in formats that resonate with state/territorial archive staff, partner associations, and stakeholder groups.

Distribute documents widely in state government archival community and to as many partner associations as possible.
Testing

Apples to apples in archival environments

- Common datasets
- Common amount of time
- Common amounts of staff intervention

"Two Apples" by lluisribesmateu1969 is licensed under CC BY-NC 2.0
Mentoring and Assistance

Mentor relationships will further support additional growth and strengthen ties.

CoSA will provide direct support to select state/territorial archives to develop the foundation needed to preserve state email records.

"Supporting cast" by FotoFloridian is licensed under CC BY-NC 2.0
Community

Making the community larger

Expanding what the community works together on
Email Archives: Building Capacity and Community

“Email Archives: Building Capacity and Community” is a four-year program seeking to build email archiving capacity in archives, libraries, and museums. The program will fund projects of $25,000 to no more than $100,000 dollars. The goal of this project is to build a broad community of institutions that can preserve email as part of their research collections. We invite you to be a part of this important initiative by submitting a proposal!
Project Ideas

• Efforts that implement ‘low barrier’ email processing workflows ([https://www.ideals.illinois.edu/handle/2142/109251](https://www.ideals.illinois.edu/handle/2142/109251))
• Tools to harvest linked content
• Projects focusing on local government emails and non-government organizations
• Community email archiving
• Your ideas!
Thank you!

Ruby Martinez - rubylm2@illinois.edu
Matthew Connelly - mjc96@columbia.edu
Gregory Wiedeman - gwiedeman@albany.edu
Michelle Gallinger - mgallinger@gallingerconsult.com