PRESERVING EMAIL IN MULTIPLE FORMATS:
AN OVERVIEW OF THE MAILBAG PROJECT

Use the chat box to tell us who you are, where you're from, and who is participating with you today.

(To open the chat window, click on the Chat icon in bottom of the screen.)

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

WELCOME!

Bonnie Weddle
New York State Archives

Fall 2021 SERI webinars

- October 8: Fantastic Bits and How to Preserve Them: A SERI Webinar for Electronic Records Day
- October 12: Advocacy and Electronic Records

In the CoSA Resource Center

- 150+ resources re: "email"
  - https://www.statearchivists.org/research-resources/resource-center

CoSA PREPARE (Preparing Archives for Records in Email)

- https://www.youtube.com/watch?v=e3hKBLLTi0k

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

SERI 2020-21 VIDEO SERIES

- **Introduction**
  - https://www.youtube.com/watch?v=OwLx8-E-l8M&list=PLospvJp0HcS9HJHf1e-rYAuk2KJFt5gV5&index=13
- **Key Concepts in Digital Processing**
  - https://www.youtube.com/watch?v=kYd7sG_3aw8&list=PLospvJp0HcS9HJHf1e-rYAuk2KJFt5gV5&index=12
- **Making the Pitch for Electronic Records**
  - https://www.youtube.com/watch?v=35xLibzrIew&list=PLospvJp0HcS9HJHf1e-rYAuk2KJFt5gV5&index=6
- **Long-Term Preservation: File Formats**

⑩ https://www.youtube.com/watch?v=eJfnuSo_lCQ&list=PLospvJp0HcS9HJHf1e-rYAuk2KJFt5gV5&index=2

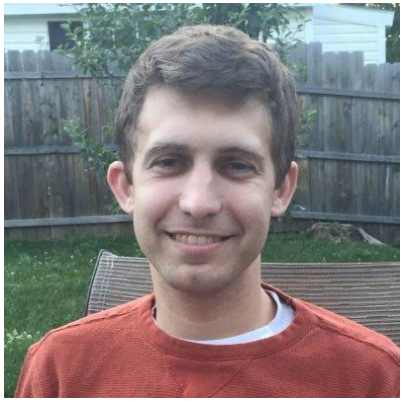
Additional videos coming soon!

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

SERI 2021 SPONSOR

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

TODAY'S PRESENTER



Greg Wiedeman

University Archivist

University at Albany, SUNY

Mailbag Project PI

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG PROJECT

- Currently there is no single effective preservation format for email
- Specification & open source tool
- Preserves email archives over IMAP or common export formats
- Enable basic near-to-capture processing
- Mailbag Website: https://archives.albany.edu/mailbag/
- Mailbag GitHub: https://github.com/UAlbanyArchives/mailbag
- Slides: https://archives.albany.edu/patron/SERI_Mailbag_webinar.pptx

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

WEBINAR OVERVIEW

1. Where we're at with email archives
2. Project origin
3. Preservation issues with email
4. The Mailbag niche
5. Grant design/process
6. Specification development
7. Community feedback process
8. Tool development
9. Q&A

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

WHERE WE ARE AT/SELF-ASSESSMENT

PollEv.com/gregorywiedeman457

Text GREGORYWIEDEMAN457 to 37607

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

PROJECT ORIGIN

- How can we document elections in New York?
- 2016 pilot project to collect fundraising emails from federal-level incumbent candidates
- 2018 expanded to state Senate and Assembly, Governor's race
- Signed up for email on candidate websites
- Used Gmail account
- MBOX file export added to repository workflow as SIP

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

PROJECT ORIGIN

- **I have an MBOX file, now what?**
  - Tried processing in early 2019
  - Python scripts to extract HTML
  - Convert to PDF with wkhtmltopdf

PROJECT ORIGIN

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 >>**

THE HILL

Just In...

Seven rising Democratic stars

BY CRISTINA MARCOS - 01/02/17 04:02 PM EST

Joe Crowley (N.Y.)

Crowley was already a shoo-in to become the new Democratic caucus chairman when colleagues started urging him to mount a challenge to Pelosi.

The 54-year-old Queens congressman is popular among fellow Democrats and is a prolific fundraiser. And he's hard to miss in a crowd with his deep, booming voice and 6-foot-5-inch stature.

While Crowley passed on running for minority leader this time, serving as caucus chairman will give him a high-profile role crafting House Democrats' strategy during the Trump administration.

If Pelosi, Hoyer or Clyburn move on in the coming years, Crowley would be well-positioned to offer himself as a next-generation leader with years

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

PROJECT ORIGIN

Date: Mon, 7 Nov 2016 20:05:46 +0000 (GMT)
From: Chuck Schumer
Reply-To: info@chuckschumer.com
To: ualbanymodernpoliticalarchives@gmail.com
Message-ID: <1175757182.26537643147854914602.JavaMail.app@rbg23.atlis1>
Subject: 30 hours

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.

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

PRESERVING EMAIL

- **Email exports don’t contain all the information**
  - Images, CSS can be hosted on external servers
  - Email CSS can be interactive!
  - Links to other web content
  - Email marketing software obfuscates URLs
- **MBOX file was rapidly degrading!**

```
<a
href="http://click.actionnetwork.org/mpss/c/1AA/ni0YAA/t.23p/WmZ7-VSKRVyi4FKQ6PoSHg/h0/kilmq4yahfKGsjd
DvZO5dj7B0XturWbGjWZyESALZkPFDS-2BuNGPThd45Z-2BbqBmNsg7H5DlnAkUjAUI6n10j90N-2FwMa3Kx6Sy9-2BkOPBGnA9TQg
1HEbnVDROqrrtMa28eX1mx7EgjDCeWoO89G-2BBXbAHN9lTOUnz5YfvEw97c76tPFoA-2FA2s2jX2yXh8SRKlg2JKSfmSvMNgyANBZ
Y0vS52njKg2BEomRSFLCQBRL1BG9aHx-2FqAyjV0ApGd-2B-2FhdF3tWuRX3s3hmgE4EjxWsGwroQ-3D-3D"
style="color:blue;font-weight:bold;color:blue">Click here to read the full article >></a>
```

SERI State Electronic Records Initiative
CoSA COUNCIL OF STATE ARCHIVISTS

PRESERVING EMAIL

- **Raw email exports**
  - Open: MBOX, EML
  - Proprietary: PST, MSG
  - Preserve email sufficiently as structured data
  - Missing data/rapid decay
- **PDF files**
  - Preserves email well as visual document
  - Does not structure data well for computational use
- **WARC files**
  - Preserves visual document experience well
  - Maintains data structure
  - Hard to use/low support for email

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

LET'S USE MULTIPLE FORMATS!

- **Multiple preservation formats**
- **Different options for use cases**
  - MBOX + PDFs
  - MBOX + PDFs + WARCs
  - PST + EMLs + PDFs
  - PST + MBOX + PDFs + WARCs
  - EML + PDFs + WARCs
- **How do we maintain relationships?**
- **Bagit specification is widely used**
  - Flexible structure
  - Provides fixity
  - Bagit-python

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG NICHE

- **Common specification to provide structure for preserving email using multiple preservation formats**
  - in place of a single preservation format
- **Basic open source utility to create and manage mailbags**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG NICHE

- **Aims to solve technical challenges**
  - Preserves email so it won't degrade if actively managed
  - Makes acquisition, processing, basic access easier
- **Does not solve human challenges**
  - Records management
  - Personally identifiable information (PII)
  - General privacy concerns
  - Legal risk aversion
- **These are much harder, require humans**
- **Hopefully make email archives more visible**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG NICHE

- **Feasible project, existing proof of concept**
- **“Separation of concerns” principle**
  - Do one thing well
- **Easy to use**
- **Inclusive, allow the most possible archivists to use**
- **Mailbag will not solve all your email archives problems, but it will do a very good job at solving some email archives problems**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

GRANT DESIGN

- **Email Archives: Building Capacity and Community program**
  - University of Illinois Mellon Foundation re-grant program
  - Up to $100,000, no indirect costs/time buyouts
- **How do we build this?**
  - Paying for stuff is easy, paying people is hard
  - Awesome professional community
  - Great students at UAlbany!
- **Build sustainably**
  - Constantly keep maintenance in mind
  - No great vendor for this
  - Students have diverse experiences

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

GRANT DESIGN

- **Advisory board of community experts**
  - Hands-on experience
- **“Bake-in” learning to the work**
  - Experienced developer as consultant
    - Worked on similar projects with sustainable focus
  - Two graduate students
  - Collaborative code review process
- **Support remote conference attendance**
- **Honorariums to support community project work**
  - Specification development
  - Code reviews
  - Documentation writing
- **Build-in outreach**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

GRANT DESIGN

- **Graduate Student Developers** **$47,132**
  - Doubled library GA program rate
- **Consultant Developer, 100 hours** **$7,000**
- **Community honorariums** **$6,400**
- **Conferences** **$1,350**
- **Cloud servers** **$1,262**
- **Marketing** **$700**
- **Total:** **$63,890**
- **I would have added more consultant/honorarium time**
- **Lots time to administer**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

SPECIFICATION DEVELOPMENT

- **Open specification, independent from Mailbag tool**
  - Allow for additional implementations
  - Support broader use cases than we can resource
  - Promote interoperability
- **Honoraria to support, altered original plan**
- **Project team wrote first draft**
- **Initial feedback from advisory board**
- **2 hour specification working meeting**
  - Open call for participants
  - Collaborative exercises to gather community input
  - Great feedback, exercises needed more time

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG SPECIFICATION

```
<base directory>/
        |
        +-- bagit.txt
        |
        +-- bag-info.txt
        |
        +-- mailbag.csv
        |
        +-- manifest-<algorithm>.txt
        |
        +-- tagmanifest-<algorithm>.txt
        |
        +-- data/
            |
            +-- mbox/
            |    +-- [payload files]
            +-- pdf/
            |    +-- [payload files]
            +-- warc/
            |    +-- [payload files]
            +-- attachments/
                  +-- [Mailbag-Message-ID]/
                  |    +-- [payload files]
                  +-- [Mailbag-Message-ID]/
                  |    +-- [payload files]
                  ...
```

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG PAYLOAD

```
+-- mailbag.csv
|
+-- manifest-<algorithm>.txt
|
+-- tagmanifest-<algorithm>.txt
|
+-- data/
    |
    +-- mbox/
    |    -- All mail Including Spam and Trash.mbox
    +-- pdf/
    |    -- 8sPf2WLSpyp65KLcYNgpX5.pdf
    |    -- LVGxWjUABLVB5dep5hZTiZ.pdf
    |    -- 9SXvpbe2AtgvedwZJCfWF4.pdf
    |       ...
    +-- warc/
    |    -- 8sPf2WLSpyp65KLcYNgpX5.warc.gz
    |    -- LVGxWjUABLVB5dep5hZTiZ.warc.gz
    |    -- 9SXvpbe2AtgvedwZJCfWF4.warc.gz
    |       ...
    +-- attachments/
        +-- 8sPf2WLSpyp65KLcYNgpX5/
        |    -- Image1.jpg
        |    -- Image2.jpg
        +-- 9SXvpbe2AtgvedwZJCfWF4/
             -- original_filenames.txt
             -- packageList.pdf
             -- draft_for_review.odt
             -- 9SXvpbe2AtgvedwZJCfWF4-2.odt
```

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG PAYLOAD

```
|
+-- data/
    |
    +-- mbox/
    |    +-- All mail Including Spam and Trash.mbox
    +-- eml/
    |    +-- Inbox
    |    |    +-- [payload files]
    |    +-- Sent Mail
    |    |    +-- [payload files]
    |    +-- Junk
    |         +-- [payload files]
    +-- pdf/
    |    +-- Inbox
    |    |    +-- [payload files]
    |    +-- Sent Mail
    |    |    +-- [payload files]
    |    +-- Junk
    |         +-- [payload files]
    +-- warc/
         +-- Inbox
         |    +-- [payload files]
         +-- Sent Mail
         |    +-- [payload files]
         +-- Junk
              +-- [payload files]
```

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG FILE NAMING

```
|
+-- data/
    |
    +-- eml/
    |    -- A free and open internet.eml
    |    -- A simple pledge.eml
    |    -- A year of organizing.eml
    |      ...
    +-- pdf/
    |    -- A free and open internet.pdf
    |    -- A simple pledge.pdf
    |    -- A year of organizing.pdf
    |      ...
    +-- warc/
    |    -- A free and open internet.warc
    |    -- A simple pledge.warc
    |    -- A year of organizing.warc
    |      ...
    +-- attachments/
          +-- 8sPf2WLSpyp65KLcYNgpX5/
               +-- Image1.jpg
               +-- Image2.jpg
          +-- 9SXvpbe2AtgvedwZJCfWF4/
               +-- packageList.pdf
```

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

BAG-INFO.TXT

Example 2 (IMAP source and MBOX & EML & PDF & WARC derivatives):

Bag-Type: Mailbag
Mailbag-Source: IMAP
Original-Included: False
Package-Date: 2021-05-04T18:16:58+00:00
Bagging-Date: 2021-05-04
External-Identifier: 944efc3e-d9df-40ad-8b87-fbb120241ddb
Mailbag-Agent: mailbag
Mailbag-Agent-Version: 0.0.1
Capture-Date: 2021-05-04T18:16:23+00:00
Capture-Agent: imaplib
MBOX-Format-Details: MBOXO
MBOX-Software-Agent: mailbox
MBOX-Software-Version: 0.4
EML-Format-Details: EML
EML-Software-Agent: email
EML-Software-Version: 4.0.2
PDF-Format-Details: PDF
PDF-Software-Agent: pyPdf
PDF-Software-Version: 1.13
WARC-Format-Details: WARC 1.1
WARC-Software-Agent: warcio
WARC-Software-Version: 1.7.4

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG.CSV

- Mailbag-Message-ID (required)
- Message-ID (required)
- Message-Path (required)
- Original-Filename (required)
- Attachments (required)
- Date
- From
- To
- Cc
- Bcc
- Subject
- Content-Type

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

NOT RETAINED

5.4.1 Example messages_not_retained.txt

a6e213b376f34b4f839059e4cc19f2f8@jerrynadler.com
1709911313.283722191466906290194.JavaMail.app@rbg23.atlis1
c2617427190b4d0a9d45e8f741d92041@carolynmaloney.com
20160623144948.9714F608DE@233elwb03.blackmesh.com
9656357.20160623150335.576bfa478e89a7.63874426@mail1.mcsignup.com
...

5.4.2 Example folders_not_retained.txt

Drafts
Listservs/A&A
Listservs/ERS
Trash
...

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

IDENTIFIER CHALLENGES

- **Message-ID**

5aa1201370ea4e8580ca983e77197396@carolynmaloney.com

730de3b1cdef0ab5393a63ed11d02514@bounce.bluestatedigital.com

2751071166.-308783722@democracy.dsccdb.www.democratsenators.org

- **Mailbag-Message-ID**
  - 1, 2, 3
  - 8sPf2WLSpyp65KLcYNgpX5
- **Original filenames for EML, MSG, etc.**
  - “A year of organizing.eml”

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

COMMUNITY FEEDBACK

- **Open call for feedback on design documents**
  - Personas
  - Use cases
  - Requirements
- **Open call for feedback on specification**
  - Google doc
  - Github issues
  - Highlighted major feedback from working meeting

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

COMMUNITY FEEDBACK

- **Multiple export files**
- **Multiple versions**
  - Redactions
  - Weeding
- **Descriptive metadata**
  - A Mailbag is not designed to be a Submission Information Package (SIP)
  - A very different problem
  - Will support any custom bag-info.txt fields and additional tag files as bagit does

MULTIPLE EMAIL VERSIONS

- Redaction and weeding are common if not ubiquitous for email archives
- There are a wide variety of workflows
- We would only be able to support a few kinds of simple workflows
- A different type of problem
- By doing less, we can make Mailbag more useful
- Multiple mailbags, flexibility in what formats are included
- Keeping access copies together with originals may not be great practice

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MULTIPLE EXPORT FILES

- **Use cases:**
  - I have a group of export files from multiple email accounts
  - I have a group of export files from the same account, exported over time
  - I have a group of export files separated for purely technical reasons, such as file size limitations

MULTIPLE EXPORT FILES

- **Really challenging to model**
  - One-to-many becomes many-to-many
  - Duplicate messages
  - Relationships between export files
  - Introduces a lot of complexity
- **Limited resources to develop and maintain**
- **Little technical costs to multiple mailbags**
  - Can put multiple mailbags in a bag or SIP
  - Can manage mailbags as ZIP or TAR
- **This is a workflow or file management problem, not really a email preservation problem**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

THINGS I'M THINKING ABOUT

- **Archivists don't have good tools to manage born-digital files**
- **Archivists may have to rely on ordinary desktop tools to manage these challenges**
  - Ordinary tools are better resourced
  - We need more tools that give archivists this power/control
  - Desktop filesystems over web applications

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

TOOL DEVELOPMENT

- **Consultant developer outlined structure Summer 2021**
- **GA developers started end of August 2021**
- **Using GitHub tools**
  - Issues
  - Pull requests
  - Project

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

TOOL DEVELOPMENT

- Features described as issues using a template
- Assigned to one GA developer
- Code reviewed by second GA developer
- Collaborative learning
- Some contributions reviewed by Consultant developer or community member funded by honorarium

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

NEXT STEPS

- **Continue tool development**
- **Real world test for specification**
- **Mailbag Specification release by end of 2021**
- **Outreach and minimally working mailbag tool in Spring 2022**
  - Possible datathon/workshop(s)
- **Mailbag tool release near end of 2021-2022 academic calendar**

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG LINKS

- Mailbag Project Website
- Mailbag draft specification
- Mailbag GitHub

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

Q&A AND THANKS!

Project Team

- Gregory Wiedeman, University at Albany, SUNY
- Mark Wolfe, University at Albany, SUNY
- Karen Kiorpes, University at Albany, SUNY
- Harit Garg, University at Albany, SUNY
- Baibhav Rajbhandari, University at Albany, SUNY

Advisory Board

- Rachel Appel, University of Pennsylvania
- Hillel Arnold, Rockefeller Archive Center
- Mat Kelly, Drexel University
- Albert Rozo, Penn State University
- Nathan Tallman, Penn State University
- Bonnie Weddle, New York State Archives

Consultant Developer

- Dave Mayo

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

MAILBAG FEEDBACK

PollEv.com/gregorywiedeman457

Text GREGORYWIEDEMAN457 to 37607

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS

STAY CONNECTED & INFORMED

CoSA Website
http://www.statearchivists.org

CoSA Twitter Handle
@StateArchivists

CoSA Facebook Page
www.facebook.com/CouncilOfStateArchivists

CoSA SERI State Electronic Records Initiative COUNCIL OF STATE ARCHIVISTS