Task Force on Technical Approaches to Email Archives: Project Update

Chris Prom, University of Illinois at Urbana-Champaign
prom@illinois.edu

Kate Murray, Library of Congress
kmur@loc.gov
Roadmap

- Task Force Background and Workplan
- Review Email as a Technical Problem
  - Researcher Needs
  - What is Email?
  - Email Lifecycle Approach (understanding workflows and repository needs)
- Review Technical Challenges
- Describe Some Tools and Approaches We Are Assessing
- Next Steps and Questions
Sponsorship

THE ANDREW W. MELLON FOUNDATION

Digital Preservation Coalition
Charge

- (a) reexamine and assess current efforts to preserve email;
- (b) articulate a conceptual and technical framework in which these efforts can operate not as competing solutions, but as elements of an interoperable toolkit to be applied as needed;
- (c) construct a working agenda for the community to construct this technical framework, adjust existing tools to work within this framework, and begin to fill in missing elements.

http://www.emailarchivestaskforce.org
Membership

Executive Committee
Kate Murray (co-chair), Library of Congress
Christopher Prom (co-chair), University of Illinois at U-C
Fran Baker, University of Manchester
Matthew Connelly, Columbia University
Wendy Gogel, Harvard Library Preservation Services

Hillel Arnold, Rockefeller Archive Center
Courtney Cain, University of Illinois at U-C
Euan Cochrane, Yale University Library

Kevin DeVorsey, NARA
Glynn Edwards, Stanford University Libraries
Riccardo Ferrante, Smithsonian Institution Archives
William Kilbride, Digital Preservation Coalition
Jessica Meyerson, The University of Texas at Austin
Erin O’Meara, University of Arizona Libraries
Michael Shallcross, University of Michigan
Joel Simpson, Artefactual Systems
Richard Whitt, Google
Julian Zbogar-Smith, Microsoft
Friends and Consultation Pathways

- National Historical Publications and Records Commission (NHPRC)
- Council of State Archivists/NAGARA (Webinars)
- Museums and Web (April 22 in person)
- North Carolina Department of Natural and Cultural Resources
- Digital Preservation Coalition: UK Briefing Days (July and October)
<table>
<thead>
<tr>
<th>Date</th>
<th>Task</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 2016</td>
<td>Launch meeting: Identify problem</td>
<td>Completed</td>
</tr>
<tr>
<td>November 2016 through March 2017</td>
<td>Working groups meet by phone/email to collaboratively draft text and formulate provisional findings and recommendations</td>
<td>Completed</td>
</tr>
<tr>
<td>April 2017</td>
<td>Stakeholder consultations including CNI, COSA (April 18), NAGARA, Museums and Web, Open webinar</td>
<td>In process</td>
</tr>
<tr>
<td>May 2017 through August 2017</td>
<td>Working groups incorporate feedback and continue drafting Co-chairs and assistant edit draft consultation report</td>
<td>To come</td>
</tr>
<tr>
<td>July 6 2017</td>
<td>DPC Briefing to get UK input on early drafts; co-chairs in UK (in person)</td>
<td>To come</td>
</tr>
<tr>
<td>August through October 2017</td>
<td>Open feedback solicited and incorporated into the report,</td>
<td>To come</td>
</tr>
<tr>
<td>September - Oct 2017</td>
<td>Executive Committee Meeting, DPC Briefing in UK</td>
<td>To come</td>
</tr>
<tr>
<td>November</td>
<td>Full task Force meeting in New York to conclude work and approve report</td>
<td>To come</td>
</tr>
<tr>
<td>December 2017</td>
<td>Task Force Report Published</td>
<td>To come</td>
</tr>
</tbody>
</table>
Email as Cultural Documentation

Miriam Blacks Scrapbook, University of Illinois Archives

archives.library.illinois.edu/ebert/a-student-writer-to-reckon-with/ars-gratia/
Email and Iran Contra

http://nsarchive.gwu.edu/NSAEBB/NSAEBB113/north07.pdf

John Poindexter

historycommons.org
Email Matters
Email as a Habitat

- Email created within an organisation as part of its business, which may be transferred to an in-house archive at the end of its active life.
- Email forming part of archives created by individuals or organisations which are subsequently acquired by libraries or archives (generally referred to as ‘collecting institutions’) from their original creators.
- Email which is created and preserved by private individuals as part of their own, or their family’s, digital archive.
- Email created and curated by informal groups or societies outside of a corporate recordkeeping structure.
Carl Woese

● Evolutionary Microbiologist at the University of Illinois
● Revolutionary methods to understand characteristics of organisms--conservation of ribosomal DNA fingerprints
● Discovered the third domain of life: Archaea
Email is both Record and Cultural Documentation

Jorge Chapa Email Series
Date Range: January 1, 1960 to October 10, 2016
Messages: 30553 incoming, 1934 outgoing.

- Correspondents (6887)
- Persons (39050)
- Organizations (55597)
- Locations (23773)
- Image attachments (3732)
- Document attachments (3577)
Email is a Many-Splendored Thing

The term “email” is used to describe:

- An individual electronic mail message
- All of the messages ever sent and received by an account
- The act of composing, sending or reading messages
- Applications and systems that are used to transport, receive, and manage messages

Personal information manager (PIM) tools include email, internet chat, calendars, contacts and task all in one application
Diagram 1: Types of Email Data
Standardization and Interoperability

Create Message

Email Client (MUA)
Compose message using IMF/RFC 5322 for header and body;
Packages as PST, EML, MBOX, MSG, etc

Transfer Message

Email Server/Message Transfer Agent (MTA)
Transfer of email messages from one server to another; Specified by SMTP, RFC 5321: Ex: Microsoft Exchange, Postfix.

Optional: Email Server/Message Transfer Agent (MTA) 2

Optional: Message Delivery Agent (MDA)

Delivery Message

Read/Store Message

Email Client (MUA)
Compose message using IMF/RFC 5322 for header and body;
Packages as PST, EML, MBOX, MSG, etc
DCC Curation Lifecycle Model

http://www.dcc.ac.uk/resources/curation-lifecycle-model
Email Lifecycle

Message Creation
- Email client dictates format, metadata
- RM rules might apply

Message Transit
- Magic of IMF and MTA - message transfer agent which makes hard work look easy

Message Received
- Email client dictates format
- RM rules might apply

Email in Active Use
- Communication
- Reminders
- Record keeping of thoughts, actions, appointments, passwords,

Email as Record
- Self Appraisal
- Self Selection
- Archival Appraisal
- Archival Selection

Email in Archive
- Disposition
- Transfer and acquisition
- Processing
- Ingest

Email in Research
- Discovery
- Access, use and reuse
Technical Challenges

- Capture
- Ensuring Authenticity
- Working at Scale
- Addressing Security Concerns
- Messages and Formats
- Preserving Embedded Attachments and linked documents
- Tool Interoperability

Student at Lincoln Hall in Wheelchair, ca. 1957-1967  Division of Rehabilitation Services: Director's Office Photograph File, University of Illinois Archives
Capture

- Building trust
- Aggregating data
- Workflows
- Chaining tools

Credit: Digital Preservation Business Case Toolkit
http://wiki.dpconline.org/
Ensuring Authenticity

- Ingest methods matter
- Key functions and metadata stored in systems, not outputs
- Preserving the chain of actions that archivists take.
Hi Chris and Kate:

Here is a link to the draft of what I have so far for the report:

https://docs.google.com/document/d/1xrA1RmFq57sGjisf0HUibMgda_m8q9-H5vnNCxpVjA/edit?usp=sharing

It just covers what's outlined below so far, but I'll try to make more progress tomorrow on the rest of it. Can you remind me of how many words you would want my section to be?

Best,
Matt

Topics:
Working at Scale

- Use and number of accounts
- Size of account vis a vis resources to preserve
- Exmamp
Specific vulnerabilities include malicious content, spoofing, confidentiality, Third Party Privacy.
Discovery and End User Access

- In room?
- On web?
- Redacted?
- Aggregate Description?
- Item level pathways?
Accessioning, Processing and Interoperability

- Significant properties?
- Semi-standardized workflows?
- API development?
- User experience?

Digital Preservation Business Case Toolkit
http://wiki.dpconline.org/
Preservation Approaches: Format Migration

- Many email tools are format dependent
  - ePADD ingests only MBOX or IMAP accounts
  - Harvard EAS ingests many formats but normalizes to EML for processing and preservation
- Migration is always a risk vs reward conversation:
  - Some formats play better together than others
  - Open non-proprietary formats are better than closed, proprietary formats
- Migration for MESSAGE format only - typically not attachments

Shameless plug for LC Sustainability of Digital Formats website:
http://loc.gov/preservation/digital/formats/index.html
Preservation Approach: Emulation

Recreating user experience for both message and attachments in original context

Software Preservation Network:
http://www.softwarepreservationnetwork.org
Preservation Approach: Bit Level Preservation

- You get what you get and you don’t get upset
- Might be appropriate for embargoed collections
Tools within Cultural Heritage Domain

- ePADD
- DArcMail
- Bitcurator
- Lots of proprietary tools such as Aid4Mail, Emailchemy, Mailstore, Access Data FTK, Preservica, and many more
- REST APIs (to come)
Tools Template

Created by Wendy M. Gogel, last modified on Mar 27, 2017

Please use this template to provide a summary of each tool:

**Tool Name**

- Basics
  - Link to information about the tool
  - Who built it?
  - Status re: availability and openness (open source, proprietary, commercial, etc.)
  - Is it actively being maintained and supported?
    - E.g.: does it have
      - active development?
      - active community?
      - services - including support?
  - System requirements

- What does the tool do?
  - 2 - 3 sentence description of purpose
  - What parts of the stewardship lifecycle does it address/support? 1-2 sentences
  - What are the tools strengths, weaknesses, and gaps?

- Ingest and Export formats
  - Which formats does the tool ingest?
  - Which formats does the tool export?
Completed Template Example

SysTools Outlook PST Viewer FREEWARE

- Basics
  - Status: Freeware version of commercial product with reduced functionality. See details below.
  - Tool is actively maintained with the most recent release 4.5 in 2017. Support is available via chat.
- What does the tool do?
  - SysTools Outlook PST Viewer FREEWARE previews PST files from MS Outlook accounts including email messages and attachments, contacts, calendars, tasks, notes and journals.
  - This tool is most helpful in early appraisal when “for fee” tools are not available.
  - What are the tools strengths, weaknesses, and gaps?
    - Because this the freeware version of a commercial tool, the functionality is reduced. Limitations include: no search capability, no mime view of message content, no export/format migration
    - Advantages include: ability to preview many attachment formats including word processing and image formats and no PST ingest file size limit.
- Ingest and Export formats
  - The tool ingests only PST files
  - No export capability in FREEWARE version.
Industry Tools and Projects

- Forensics
- Legal - email as evidence
- Email production & marketing
- Email abuse, abuse prevention & deliverability
- Consumer email services
- Enterprise email services & operations
- Email storage & data management
- Email within Records Management
State Archives Projects / NHPRC Funding

KAIN EMAIL PROJECT @ LVA

Welcome to the Library of Virginia’s Kaine Email Project, where we make accessible the email records from the administration of Governor Timothy M. Kaine, Virginia’s 70th governor (2006–2010). Users can search and view email records from the Governor's Office and his Cabinet Secretaries; learn about other public records from the Kaine Administration; go behind the scenes to see how the Library of Virginia made the email records available; and read what others are saying about the collection. The Library of Virginia received approximately 1.3 million email messages from the Kaine Administration. We are processing and releasing these records in batches, so please check back often for new content.

AutoTAR Technology-Assisted Review Platform

with Continuous Active Learning™ (CAL™)

Maura R. Grossman and Gordon V. Cormack

Search Jeb Bush Email:

hotel

* CAL and Continuous Active Learning are Trademarks of Maura R. Grossman and Gordon V. Cormack

For more information on CAL, see our Practical Law Journal Article

Also available: Search Tim Kaine Email
Harvard Electronic Archiving System - EAS

http://library.harvard.edu/preservation/email-archiving

_**Harvard Library Resources - Reports, Articles, Presentations, etc.**_

**Email Archiving Systems Interoperability** (Report, July 2016)

**Email Archiving Stewardship Tools Workshop** (Report, July 2016)

**Email Archiving Stewardship Tools Workshop** (Article, March 2016)

**Email Archiving in a Curation Setting** (Panel Presentation, March 2016)

**Reshaping the Repository: The Challenge of Email Archiving** (Paper, 2010)
Deliverables: Website Resources

In late 2016, the Andrew W. Mellon Foundation and the Digital Preservation Coalition announced the formation of the Task Force on Technical Approaches to Email Archives. The task force is charged with assessing current frameworks, tools, and approaches being taken toward email as critical historical sources.

Through a set of working groups, the task force will construct a working agenda for the community by focusing on three issues: (1) articulating this technical framework, (2) suggesting how existing tools fit within this framework, and (3) beginning to identify any missing elements. A final report is expected by December 2017.

http://www.emailarchives.taskforce.org
Bibliography


Abstract: This website, compiled by members of the American Archivists’ Science, Technology, and Health Care Roundtable (STHC) and the Archivists and Librarians in the History of Health Sciences (ALHHS), provides information on how the Health Insurance Portability and Accountability Act of 1996 (HIPPA) impacts historical research in libraries, archives, and other record repositories.


Abstract: These pages are designed to be a complete, correct, comprehensible reference for the format of an internet mail message header. They explain what shows up in today's Internet mail messages, what today's mail-reading programs can handle, and what the IETF header-format specifications say.


Abstract: Information and covered entity chart to help industry understand if they are a covered entity.
**Deliverables:**

**Draft and Final Versions of Report, including Tools Lists**

<table>
<thead>
<tr>
<th>Date Range</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2017 through April 2017</td>
<td>Working groups continue drafting report sections</td>
</tr>
<tr>
<td>April 2017</td>
<td>Task Force consults with stakeholders including COSA, NAGARA, CNI (in person April 3/4), Museums and Web (in person April 22). Review work to date and get feedback.</td>
</tr>
<tr>
<td>April 2017</td>
<td>Task Force consults DPC/UK Group. Review work to date and get feedback.</td>
</tr>
<tr>
<td>April 2017</td>
<td>Task Force Open Webinar. Review work to date and get feedback.</td>
</tr>
<tr>
<td>May 2017 through August 2017</td>
<td>Co-chairs and assistant edit report</td>
</tr>
<tr>
<td>Late June/July</td>
<td>DPC Briefing to get UK input on early drafts; co-chairs in UK (in person)</td>
</tr>
<tr>
<td>August through September 2017</td>
<td>Open feedback solicited and incorporated into the report</td>
</tr>
<tr>
<td>Early September</td>
<td>Executive Committee Meeting in New York: Review progress</td>
</tr>
<tr>
<td>October/Early November</td>
<td>DPC Briefing to get UK input on early drafts; co-chairs in UK (in person)</td>
</tr>
<tr>
<td>November</td>
<td>Full task Force meeting in New York to conclude work and approve report</td>
</tr>
<tr>
<td>December 2017</td>
<td>Task Force Report Published</td>
</tr>
</tbody>
</table>
Questions, Feedback, Discussion

Chris Prom, University of Illinois at Urbana-Champaign
prom@illinois.edu

Kate Murray, Library of Congress
kmur@loc.gov

http://www.emailarchivestaskforce.org