Recommendations from the Task Force on Technical Approaches to Email Archives: DPC Briefing Day

Chris Prom, University of Illinois at Urbana-Champaign
Kate Murray, Library of Congress

January 24, 2018
Woburn House, London
Road Map

- Brief overview of TF & Publishing Plans
- Review integrated comments from July Briefing
- Overview of Current Draft
- Focus on Recommendations - Two Main Areas:
  - Community Development and Advocacy
  - Tool Support, Testing, and Development
- Discussion this afternoon: Feedback on Recommendations and DPC Report Update
About Us: Chris Prom

- Archivist and Andrew S. G. Turyn Professor, University of Illinois
- Publications Editor, Society of American Archivists
- 2011 DPC Tech Watch Report *Preserving Email*
  - Will be working on 2nd edition later this spring
- Project Partner EPADD Project (Stanford University)
About Us: Kate Murray

- NARA’s Transfer Guidance (2014)
  - Individual messages: XML and MSG
  - Aggregated messages: PST and MBOX
- Sustainability of Digital Formats website (2014-current):
- NDSA Email Interest Group (2014-2015)
- Archiving Email Symposium (2015):
  [http://www.digitalpreservation.gov/meetings/archivingemailsymposium.html](http://www.digitalpreservation.gov/meetings/archivingemailsymposium.html)
- Harvard EAST Workshop (2016)
- LC email collections starting to come in
Charge

- (a) examine 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

Hillel Arnold, Rockefeller Archive Center
Courtney Cain, Lake Forest College
Euan Cochrane, Yale University Library

Kevin DeVorsey, National Archives and Records Administration
Glynn Edwards, Stanford University Libraries
Riccardo Ferrante, Smithsonian Institution Archives
William Kilbride, Digital Preservation Coalition
Jessica Meyerson, Educopia Institute
Erin O’Meara, University of Arizona Libraries
Michael Shallcross, University of Michigan
Joel Simpson, Artefactual Systems
Camille Tyndall Watson, State Archives of North Carolina
Richard Whitt, Google
Julian Zbogar-Smith, Microsoft
Friends and Consultation Pathways

- Council of State Archivists/NAGARA (Webinars)
- National Historical Publications and Records Commission (NHPRC)
- North Carolina Department of Natural and Cultural Resources
- Coalition for Networked Information (CNI, April 2017 and upcoming Exec Roundtable April 2018)
- Digital Preservation Coalition: UK Briefing Days

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.
Our Timeline

- Next few weeks: Incorporate feedback from this consultation and finalize draft
- Feb 28: Submit manuscript to Council on Library and Information Resources
- May: Final version of report published
Previous Comments from DPC Briefing and Other Sources

- Emphasize need to building a user community (show what can be gained).
- Challenge of “Sensitive Emails”
  - UK: GDPA
  - US: Similar Presumption of Risk
- Processing Email at Scale
- More emphasis on importance of artificial intelligence in classifying and identifying sensitive emails
- Reference research about what users want out of email collections.
More Comments from Public Review

● Report serves as both a call to arms about importance of access to email archives and exploration of technical needs to facilitate preservation
● Need for “skilling up” of workforce on digital technology literacy
● Need for sustainability plans for open source tools - move away from limited term funding to programmatic and supported solutions
Section 1: The Untapped Potential of Email Archives
Email Matters
The fact of the matter is that all the richness of email collections will remain dark until we as a community of archivists, technologists and scholars solve the technical issues inherent in its long term preservation and availability for use.

Epitaphs of the Great War:
http://www.epitaphsofthegreatwar.com/call-to-arms/
Precisely because it’s inescapable, insecure and irresistibly convenient, email provides an almost uncomfortably intimate view into the historical record. It preserves time, location and state of mind, the what-when-where-and-who of every story we might want to dig up. The last two decades, email’s high-water era, have thus been a bounty for anyone wishing to understand exactly what was happening in the inner circles of powerful organizations — for journalists, historians and prosecutors of white-collar crime, among others.
So What Needs to Change in Archival Processes?

- Embrace email as complex research data
- Harness new technology such as NLP and machine learning for actionality not easily imaginable for paper
- Encourage content creators to take an active role in preservation, even for personal papers
- Build toward greater tool interoperability and toward deeper community integration
Section 2: The Email Lifecycle
Email as ...

Organizational Record

- Documents discussions, decisions, and actions performed in the course of business
- Retention governed by legal or regulatory environment of the institution & business needs of the organization
- Disposition according to records management system & schedules
- Can be seen as liability

Personal Record/Donated Materials

- Less defined creation story, can be their own, or their family’s, digital archive, sometimes outside of a corporate record keeping structure
- Little/no organization or retention planning
- May not be considered part of larger “digital archive”
- Communication of expectations (privacy, security, access) and institutional capabilities via donor agreement is essential
Email Lifecycle

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

**Transit**
- IMF & MTA assist in moving the structured data across disparate systems

**Receipt**
- Email client dictates format, metadata
- RM rules might apply

**Active Use**
- Communication
- Documentation
- Reminders
- Actions
- Appointments
- Passwords

**As Record**
- Self appraisal
- Self selection
- Archival appraisal
- Archival selection
- RM rules might apply

**Archive**
- Disposition
- Transfer and Acquisition
- Transfer
- Ingest
- Implementation of preservation strategy

**Research**
- Discovery
- Access
- Use
Section 3:
Email as a Documentary Technology
What **IS** email?
And Why Bother with it?
Email System Architecture

- Global Addressing
- Interoperability
- Asynchronicity
- Redundancy/Best Effort Delivery
- Dispersion
- Backward Compatibility
- Extensibility
Email Message Data Model
Other elements of Email

- Accounts
- Data Transmission Model
- Security and Encryption
- “Beyond the ASCII Message”

Icons made by Freepik from www.flaticon.com and licensed CC 3.0 BY
Section 4: Current Services and Trends
Evolving Email Ecosystem: Abuse Prevention

authentication-results: spf=none (sender IP is )

x-microsoft-exchange-diagnostics:
1;BN6PR14MB1426;7: . . . .

x-ms-exchange-antispam-srfa-diagnostics:
SSOS; . . .

x-ms-office365-filtering-correlation-id: d8d3932d-150c-4d60-b8b4-08d561a63625

x-microsoft-antispam:

Phishing By Stomchak CC-BY-SA-3.0 or GFDL, via Wikimedia Commons
Compliance and Legal Discovery Tools

- Retention Management
- Journalling or ‘Archiving’
- E-Discovery

“My fees are quite high, and yet you say you have little money. I think I’m seeing a conflict of interest here.”

Image Credit: https://www.newyorker.com/cartoons/bob-mankoff/the-cartoon-lounge-grim-and-bear-it
Repository Challenges (1)

1. Capture:
   a. Direct export
   b. Web service exports
   c. Client-based exports
   d. Disk imaging

2. Ensuring
   Authenticity/Tracking
   Processing Actions

Repository Challenges (2)

4. Attachments and Linked Content

5. Security and Privacy

6. Processing at Scale: Large or Many Collections
Section Five: Potential Solutions and Sample Workflows
Preservation Approaches:
Begins with Account and Format Analysis
Preservation Approaches:

- **Bit Level:** Possibly Appropriate for Embargoed Collections
- **Format Migration**
  - Many email tools are format dependent
  - 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
- **Emulation**
  - Recreating user experience for both message and attachments in original context
  - Software Preservation Network:
    http://www.softwarepreservationnetwork.org
Tools within Cultural Heritage Domain

- Key to interoperability, scalability, preservation and access
- Several are usable and maturing--with more work coming.
EPADD: Process, Appraise, Discover, Deliver

- Stanford University
- Import via IMAP, PST and MBOX
- Export Mbox
- Entity Extraction and NLP Tools
- Discovery and Delivery environment
TOMES

● Cross-platform .pst to EAXS XML parser
● Process Capstone Accounts
● NLP dictionary flagging named entities unique to government at the state and local level
● Training Materials

https://www.ncdcr.gov/resources/records-management/tomes
- **DArcMail**
  - Smithsonian-led Project
  - Converts MBOX files to EMail Account XML Schema (EAXS).
Harvard Electronic Archiving System - EAS

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

- Aid4Mail
- Emailchemy
- Mailstore
- Access Data FTK
- Preservica

Also: Rest APIs
Workflow Scenario: Migration @ Harvard
Migration @ Stanford with ePADD
WORKFLOW SCENARIO: ACCESSING AN ENTIRE DISK IMAGE - EMAIL & ATTACHMENTS - WITH EMULATION
Section Six: The Path Forward and Next Steps
Community Development and Advocacy (nurturing and fostering)

Lower-Barrier Activities:

- Assess Institutional Readiness
- Training and Skills Development
- Demystify Email Archiving for Collection Donors
- Maintain Assessment of Email Tools in COPTR Registry
- Develop Format Comparison Matrix
Community Development and Advocacy (nurturing and fostering)

Higher-Impact Activities:

● Sustain the Email Archiving Community
● Specification Planning for Beginning-of-lifecycle Email Tools
● Develop Criteria for Email Authenticity
● Improve Standards Documentation for MBOX and EML
● Improve Options for PDF in Email Archiving Workflows
Tool Support, Testing, and Development

Lower-Barrier Activities:

- Test Existing Tools for Data Impact and Data Loss
- Improve Format Identification, Characterization, and Validation Tools for Email Formats
Tool Support, Testing, and Development

Higher-Impact Activities:

- Sustaining and Integrating Existing Tools
- Develop Email Self-archiving Tool
- Develop Standards For Tool Interoperability with a Reference Implementation
- Improved Tools for Sensitivity Review
Questions, Feedback, Discussion

Chris Prom: prom@illinois.edu
Kate Murray: kmur@loc.gov

http://www.emailarchivestaskforce.org