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): https://www.loc.gov/preservation/digital/formats/index.html
- NDSA Email Interest Group (2014-2015)
- Archiving Email Symposium (2015): http://www.digitalpreservation.gov/meetings/archivingemailsymposium.html
- Harvard EAST Workshop (2016)
- LC email collections starting to come in

Sponsorship

THE
ANDREW W.
MELLON
FOUNDATION

dpc DigitalPreservationCoalition

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

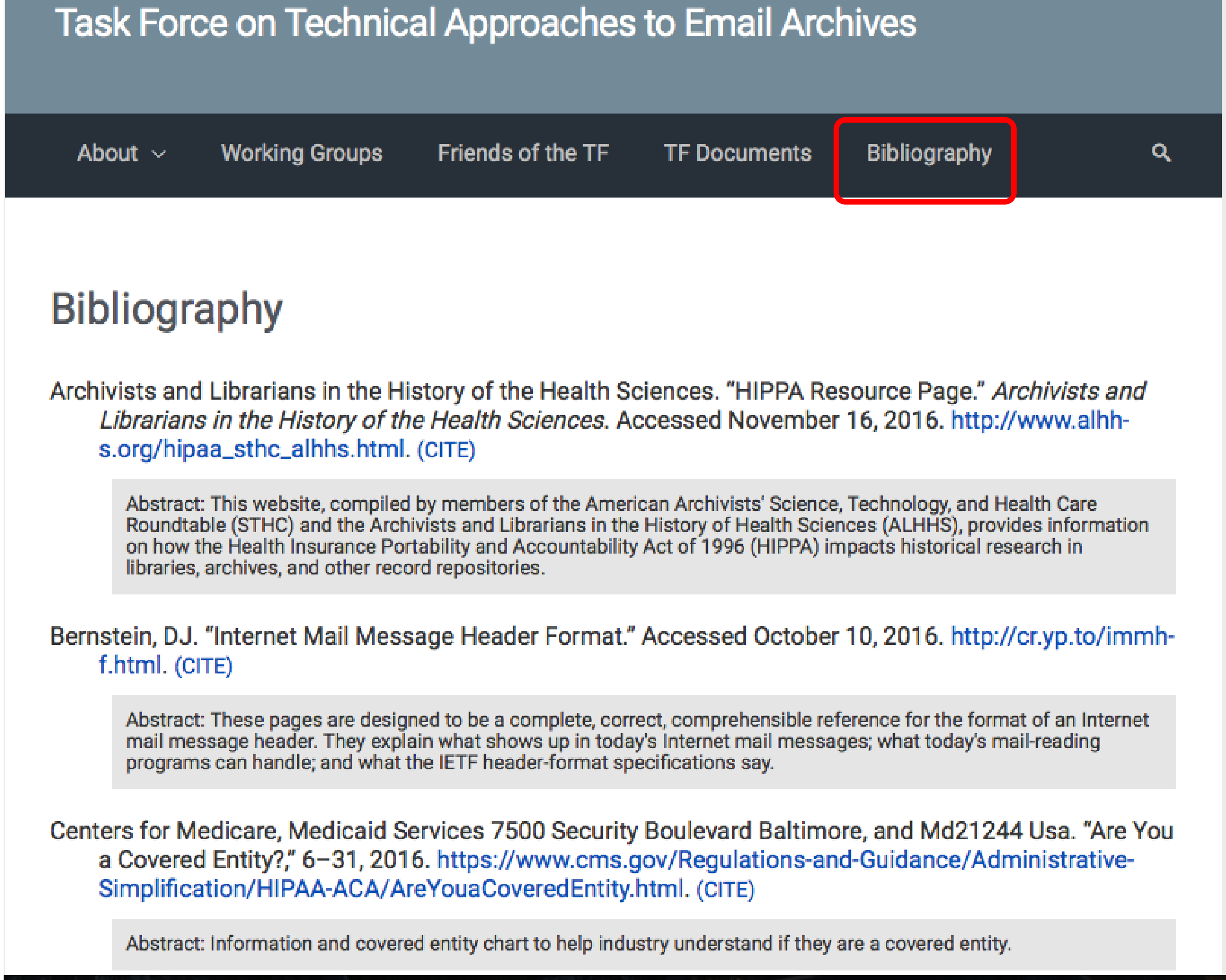
Richard Whitt, Google

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

Bibliography



Task Force on Technical Approaches to Email Archives

Bibliography

Archivists and Librarians in the History of the Health Sciences. "HIPPA Resource Page." *Archivists and Librarians in the History of the Health Sciences*. Accessed November 16, 2016. http://www.alhh-s.org/hipaa_sthc_alhhs.html. (CITE)

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.

Bernstein, DJ. "Internet Mail Message Header Format." Accessed October 10, 2016. http://cr.yp.to/immh-f.html. (CITE)

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.

Centers for Medicare, Medicaid Services 7500 Security Boulevard Baltimore, and Md21244 Usa. "Are You a Covered Entity?," 6–31, 2016. https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/HIPAA-ACA/AreYouaCoveredEntity.html. (CITE)

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

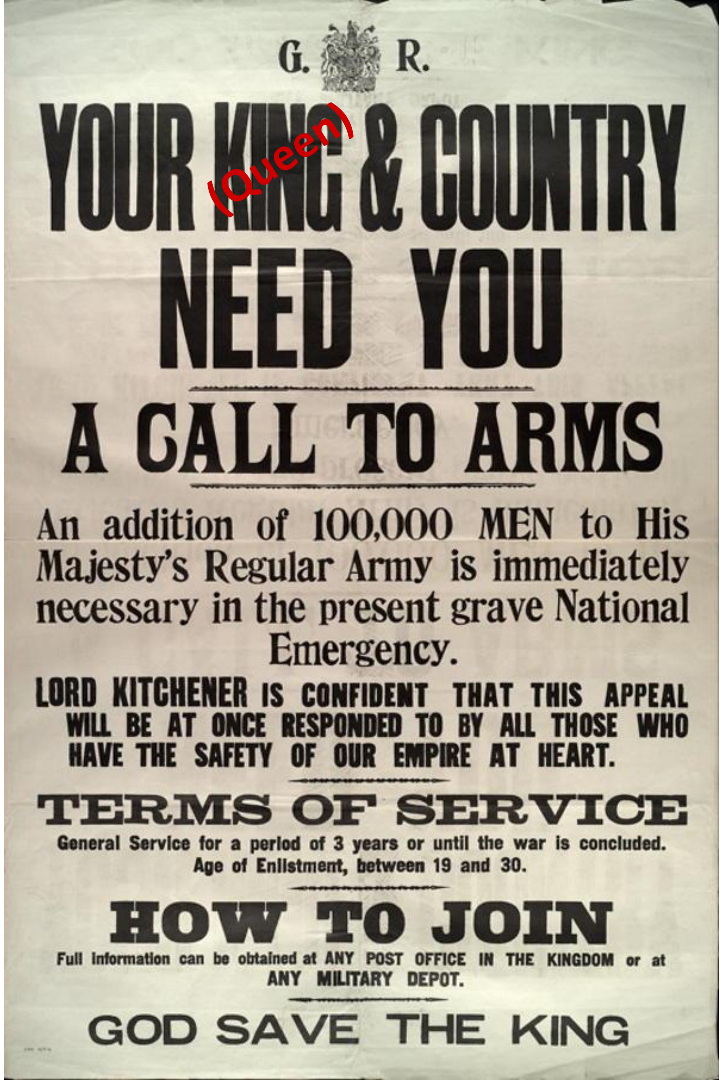
- Emphasize need to building a user community (show what can be gained).
- Challenge of “Sensitive Emails”
  - UK: GDPA
  - US: Similar Presumption of Risk
- Processing Email at Scale
- More emphasis on importance of artificial intelligence in classifying and identifying sensitive emails
- Reference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



G. R.

YOUR KING (Queen) & COUNTRY NEED YOU

A CALL TO ARMS

An addition of 100,000 MEN to His Majesty's Regular Army is immediately necessary in the present grave National Emergency.

LORD KITCHENER IS CONFIDENT THAT THIS APPEAL WILL BE AT ONCE RESPONDED TO BY ALL THOSE WHO HAVE THE SAFETY OF OUR EMPIRE AT HEART.

TERMS OF SERVICE

General Service for a period of 3 years or until the war is concluded.
Age of Enlistment, between 19 and 30.

HOW TO JOIN

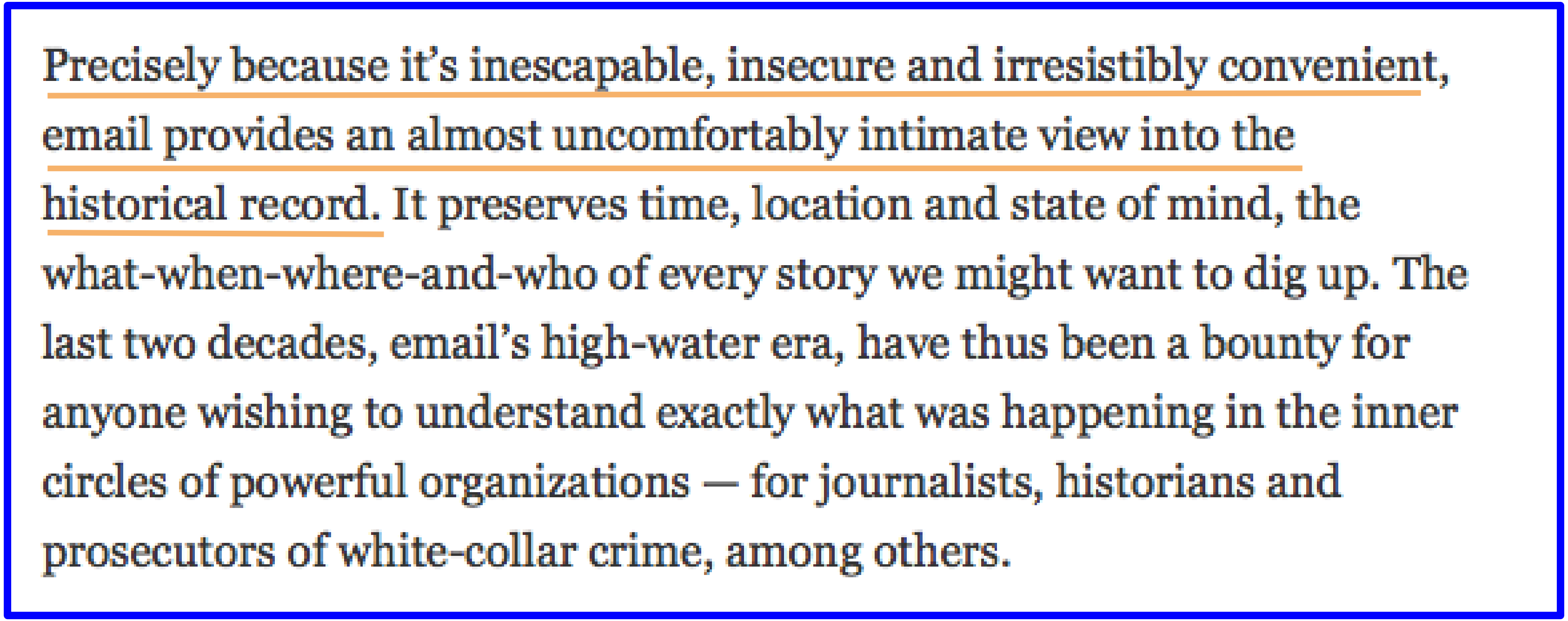
Full information can be obtained at ANY POST OFFICE IN THE KINGDOM or at ANY MILITARY DEPOT.

GOD SAVE THE KING

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/

Email is the Story Keeper



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.

NYT, What We Lose When We Lose Email, July 13, 2017

So What Needs to Change in Archival Processes?

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

Section 2: The Email Lifecycle

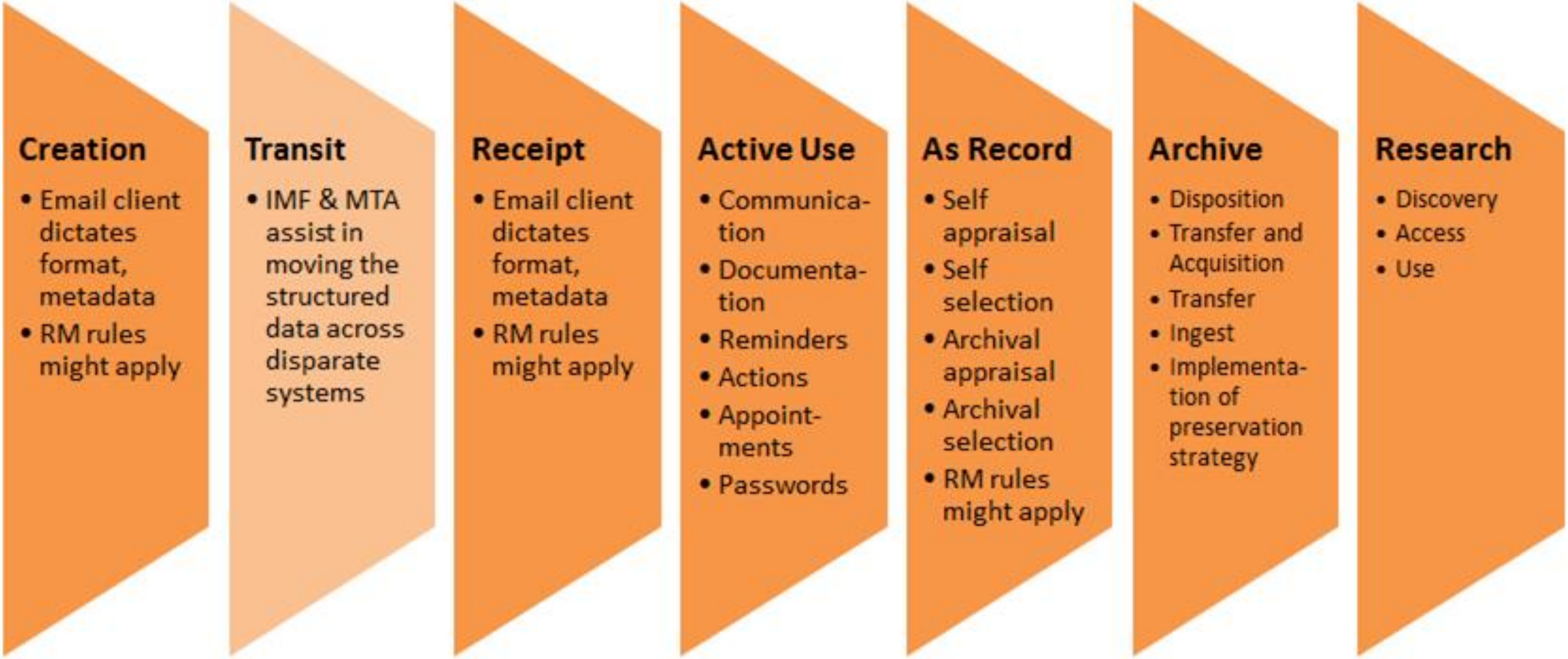
Email as ...

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



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?

CC0 Creative Commons https://pixabay.com/en/email-icon-marketing-information-1345971/

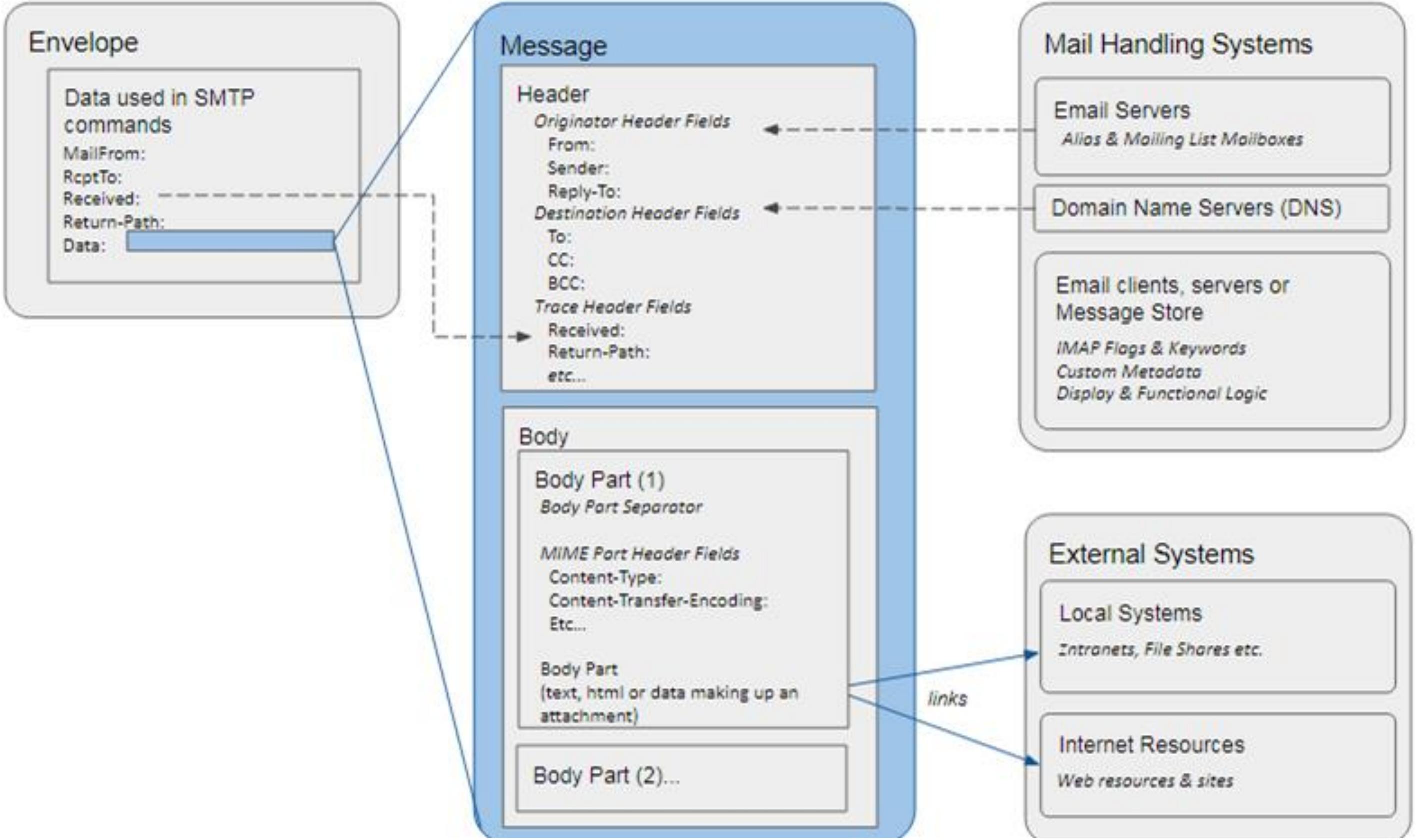
And Why Bother with it?

¯_(ツ)_/¯

Email System Architecture

- Global Addressing
- Interoperability
- Asynchronicity
- Redundancy/Best Effort Delivery
- Dispersion
- Backward Compatibility
- Extensibility

Email Message Data Model



Envelope

Data used in SMTP commands

MailFrom:

RcptTo:

Received:

Return-Path:

Data:

Message

Header

Originator Header Fields

From:

Sender:

Reply-To:

Destination Header Fields

To:

CC:

BCC:

Trace Header Fields

Received:

Return-Path:

etc...

Body

Body Part (1)

Body Part Separator

MIME Part Header Fields

Content-Type:

Content-Transfer-Encoding:

Etc...

Body Part

(text, html or data making up an attachment)

Body Part (2)...

Mail Handling Systems

Email Servers

Alias & Mailing List Mailboxes

Domain Name Servers (DNS)

Email clients, servers or Message Store

IMAP Flags & Keywords

Custom Metadata

Display & Functional Logic

External Systems

Local Systems

Intranets, File Shares etc.

Internet Resources

Web resources & sites

links

Other elements of Email

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

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

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

- Retention Management
- Journalling or 'Archiving"
- E-Discovery

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

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

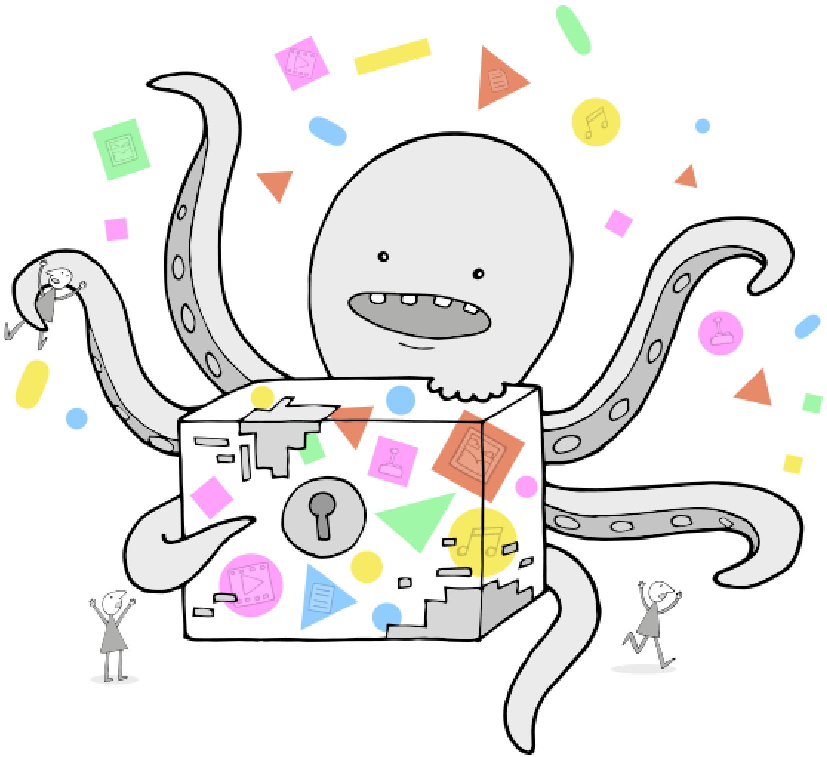
Repository Challenges (2)

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:

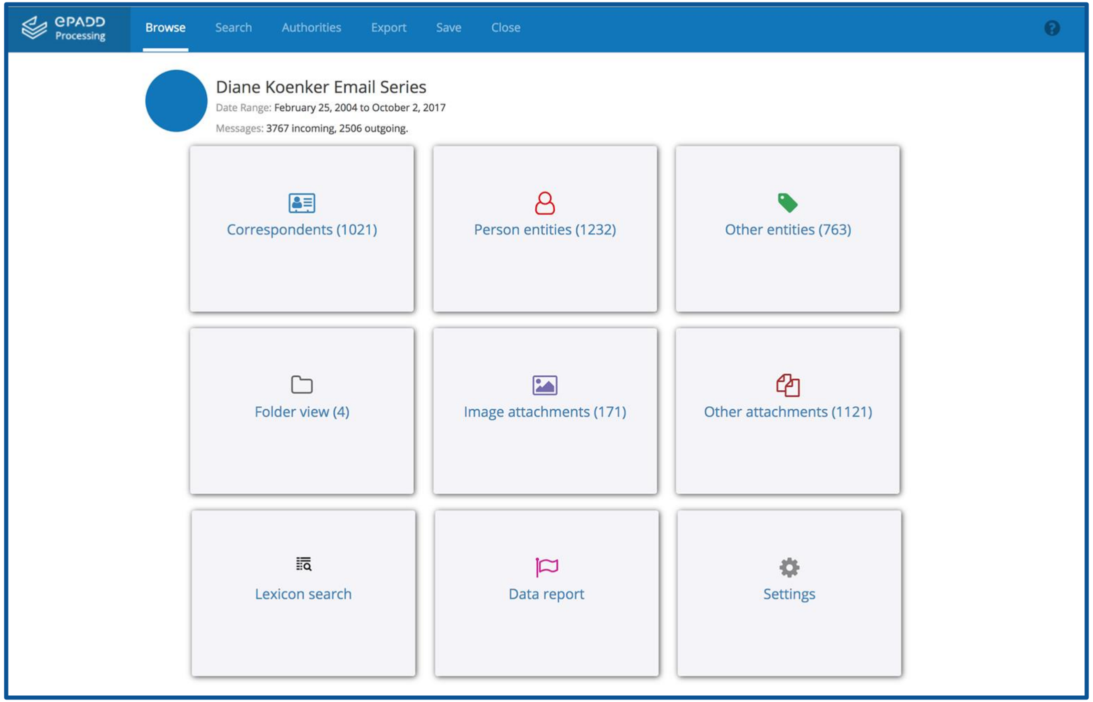
- Bit Level: Possibly Appropriate for Embargoed Collections
- Format Migration
    - Many email tools are format dependent
    - Migration is always a risk vs reward conversation:
    - Some formats play better together than others
    - Open non-proprietary formats are better than closed, proprietary 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



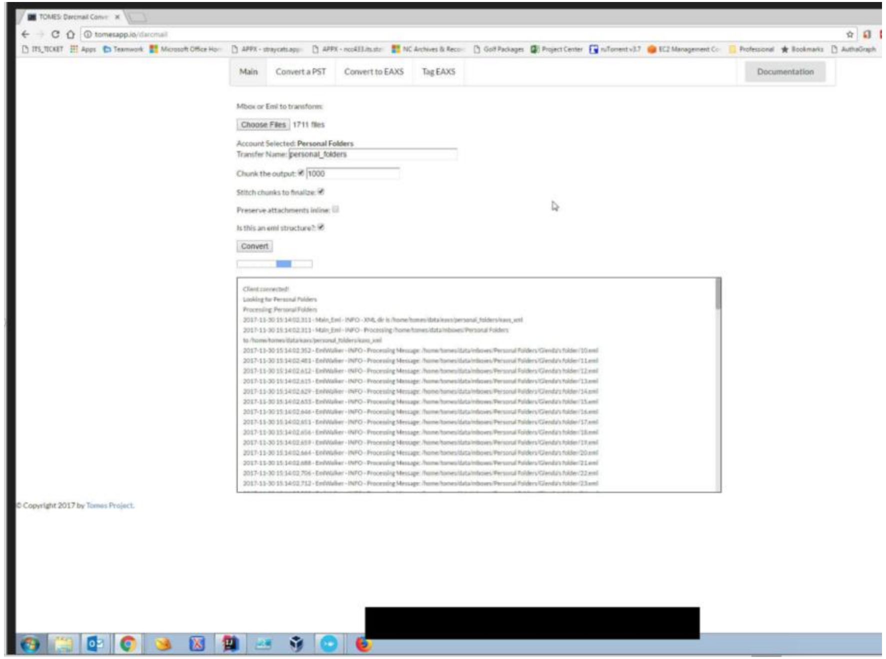
Diane Koenker Email Series

Date Range: February 25, 2004 to October 2, 2017

Messages: 3767 incoming, 2506 outgoing.

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



Mbox or Eml to transform:

Choose Files 1711 files

Account Selected: **Personal Folders**
Transfer Name: personal_folders

Chunk the output: ☑ 1000

Stitch chunks to finalize: ☑

Preserve attachments inline: □

Is this an eml structure?: ☑

Convert

```
Client connected!
Looking for Personal Folders
Processing: Personal Folders
2017-11-30 15:14:02.311 - Main_Eml - INFO - XML dir is /home/tomes/data/eaxs/personal_folders/eaxs_xml
2017-11-30 15:14:02.311 - Main_Eml - INFO - Processing /home/tomes/data/mboxes/Personal Folders
to /home/tomes/data/eaxs/personal_folders/eaxs_xml
2017-11-30 15:14:02.352 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/10.eml
2017-11-30 15:14:02.481 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/11.eml
2017-11-30 15:14:02.612 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/12.eml
2017-11-30 15:14:02.615 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/13.eml
2017-11-30 15:14:02.629 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/14.eml
2017-11-30 15:14:02.633 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/15.eml
2017-11-30 15:14:02.646 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/16.eml
2017-11-30 15:14:02.651 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/17.eml
2017-11-30 15:14:02.656 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/18.eml
2017-11-30 15:14:02.659 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/19.eml
2017-11-30 15:14:02.664 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/20.eml
2017-11-30 15:14:02.688 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/21.eml
2017-11-30 15:14:02.706 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/22.eml
2017-11-30 15:14:02.712 - EmlWalker - INFO - Processing Message: /home/tomes/data/mboxes/Personal Folders/Glenda's folder/23.eml
```

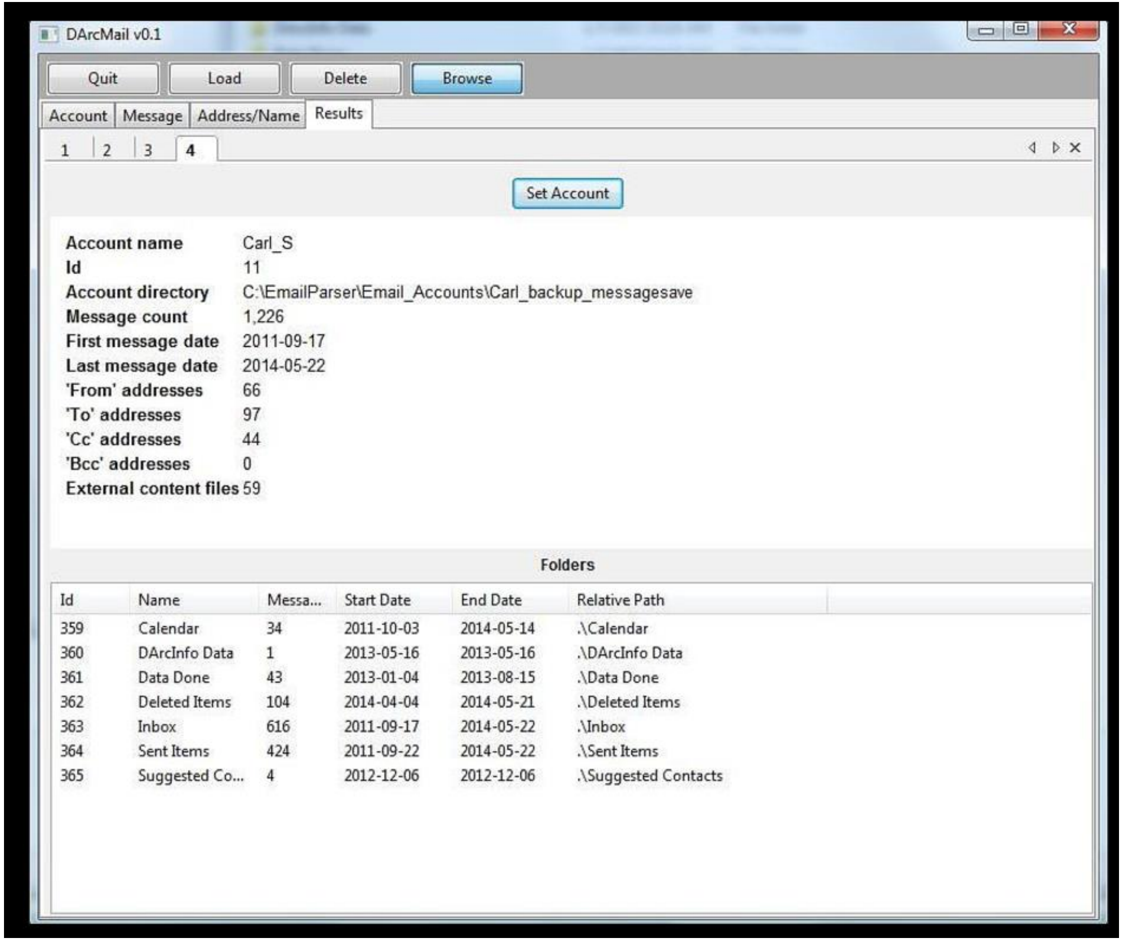
© Copyright 2017 by Tomes Project.

[redacted]

https://www.ncdcr.gov/resources/records-management/tomes

- DArcMail

- Smithsonian-led Project
- Converts MBOX files to EMail Account XML Schema (EAXS).

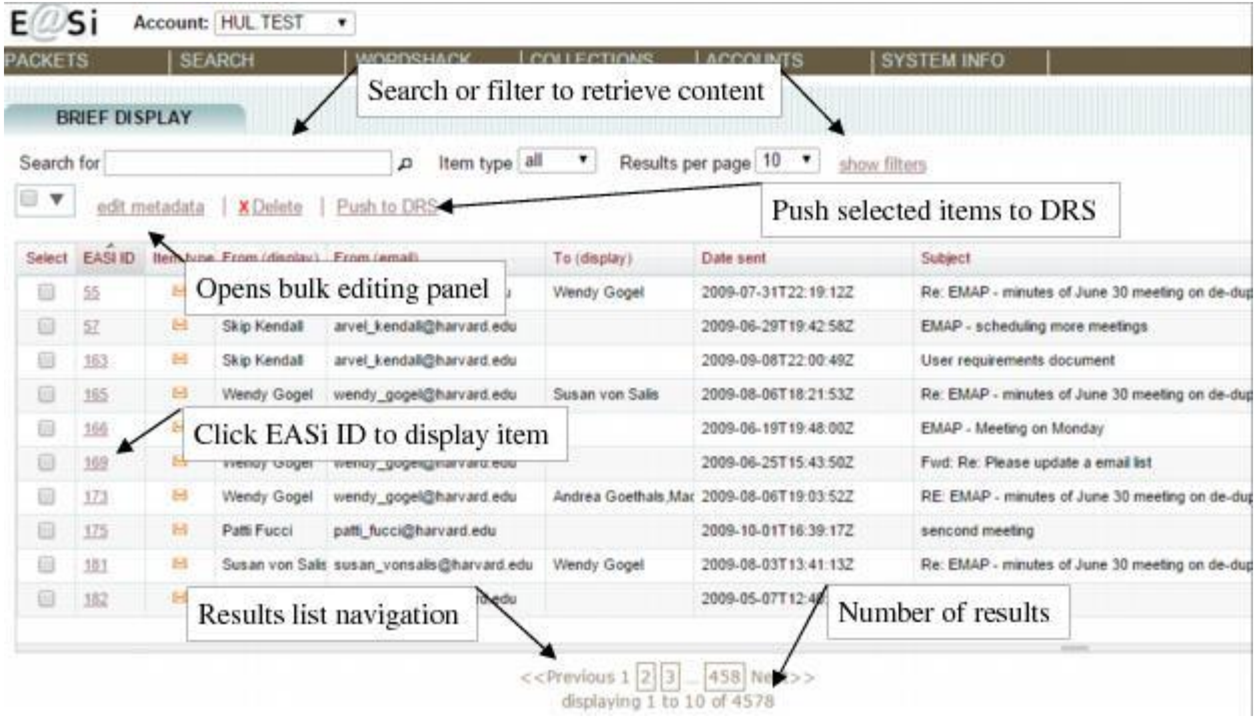


Account name	Carl_S
Id	11
Account directory	C:\EmailParser\Email_Accounts\Carl_backup_messagesave
Message count	1,226
First message date	2011-09-17
Last message date	2014-05-22
'From' addresses	66
'To' addresses	97
'Cc' addresses	44
'Bcc' addresses	0
External content files	59

Folders

Id	Name	Messa...	Start Date	End Date	Relative Path
359	Calendar	34	2011-10-03	2014-05-14	.\Calendar
360	DArcInfo Data	1	2013-05-16	2013-05-16	.\DArcInfo Data
361	Data Done	43	2013-01-04	2013-08-15	.\Data Done
362	Deleted Items	104	2014-04-04	2014-05-21	.\Deleted Items
363	Inbox	616	2011-09-17	2014-05-22	.\Inbox
364	Sent Items	424	2011-09-22	2014-05-22	.\Sent Items
365	Suggested Co...	4	2012-12-06	2012-12-06	.\Suggested Contacts

Harvard Electronic Archiving System - EAS



E@Si
Account: HUL TEST ▾
PACKETS | SEARCH | WORDSHACK | COLLECTIONS | ACCOUNTS | SYSTEM INFO
BRIEF DISPLAY
Search for
Item type all ▾
Results per page 10 ▾
show filters
□ ▾
edit metadata | ✗Delete | Push to DRS

Search or filter to retrieve content

Push selected items to DRS

Opens bulk editing panel

Click EASi ID to display item

Results list navigation

Number of results

Select	EASi ID	Item type	From (display)	From (email)	To (display)	Date sent	Subject
□	55	✉			Wendy Gogel	2009-07-31T22:19:12Z	Re: EMAP - minutes of June 30 meeting on de-dup
□	57	✉	Skip Kendall	arvel_kendall@harvard.edu		2009-06-29T19:42:58Z	EMAP - scheduling more meetings
□	163	✉	Skip Kendall	arvel_kendall@harvard.edu		2009-09-08T22:00:49Z	User requirements document
□	165	✉	Wendy Gogel	wendy_gogel@harvard.edu	Susan von Salis	2009-08-06T18:21:53Z	Re: EMAP - minutes of June 30 meeting on de-dup
□	166					2009-06-19T19:48:00Z	EMAP - Meeting on Monday
□	169	✉	Wendy Gogel	wendy_gogel@harvard.edu		2009-06-25T15:43:50Z	Fwd: Re: Please update a email list
□	173	✉	Wendy Gogel	wendy_gogel@harvard.edu	Andrea Goethals,Mac	2009-08-06T19:03:52Z	RE: EMAP - minutes of June 30 meeting on de-dup
□	175	✉	Patti Fucci	patti_fucci@harvard.edu		2009-10-01T16:39:17Z	sencond meeting
□	181	✉	Susan von Salis	susan_vonsalis@harvard.edu	Wendy Gogel	2009-08-03T13:41:13Z	Re: EMAP - minutes of June 30 meeting on de-dup
□	182	✉		d.edu		2009-05-07T12:4	

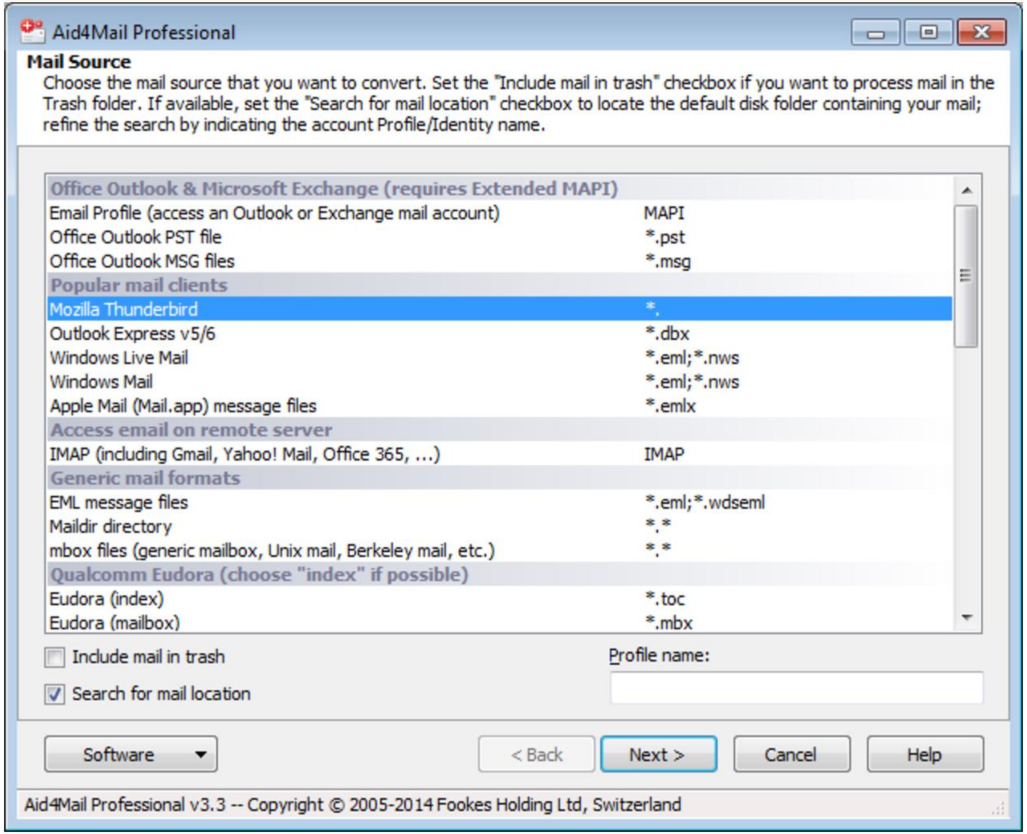
<<Previous 1 2 3 ... 458 Next>>
displaying 1 to 10 of 4578

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

Proprietary Tools

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

Also: Rest APIs



Mail Source

Choose the mail source that you want to convert. Set the "Include mail in trash" checkbox if you want to process mail in the Trash folder. If available, set the "Search for mail location" checkbox to locate the default disk folder containing your mail; refine the search by indicating the account Profile/Identity name.

Office Outlook & Microsoft Exchange (requires Extended MAPI)	
Email Profile (access an Outlook or Exchange mail account)	MAPI
Office Outlook PST file	*.pst
Office Outlook MSG files	*.msg
Popular mail clients	
Mozilla Thunderbird	*.
Outlook Express v5/6	*.dbx
Windows Live Mail	*.eml;*.nws
Windows Mail	*.eml;*.nws
Apple Mail (Mail.app) message files	*.emlx
Access email on remote server	
IMAP (including Gmail, Yahoo! Mail, Office 365, ...)	IMAP
Generic mail formats	
EML message files	*.eml;*.wdseml
Maildir directory	*.*
mbox files (generic mailbox, Unix mail, Berkeley mail, etc.)	*.*
Qualcomm Eudora (choose "index" if possible)	
Eudora (index)	*.toc
Eudora (mailbox)	*.mbx

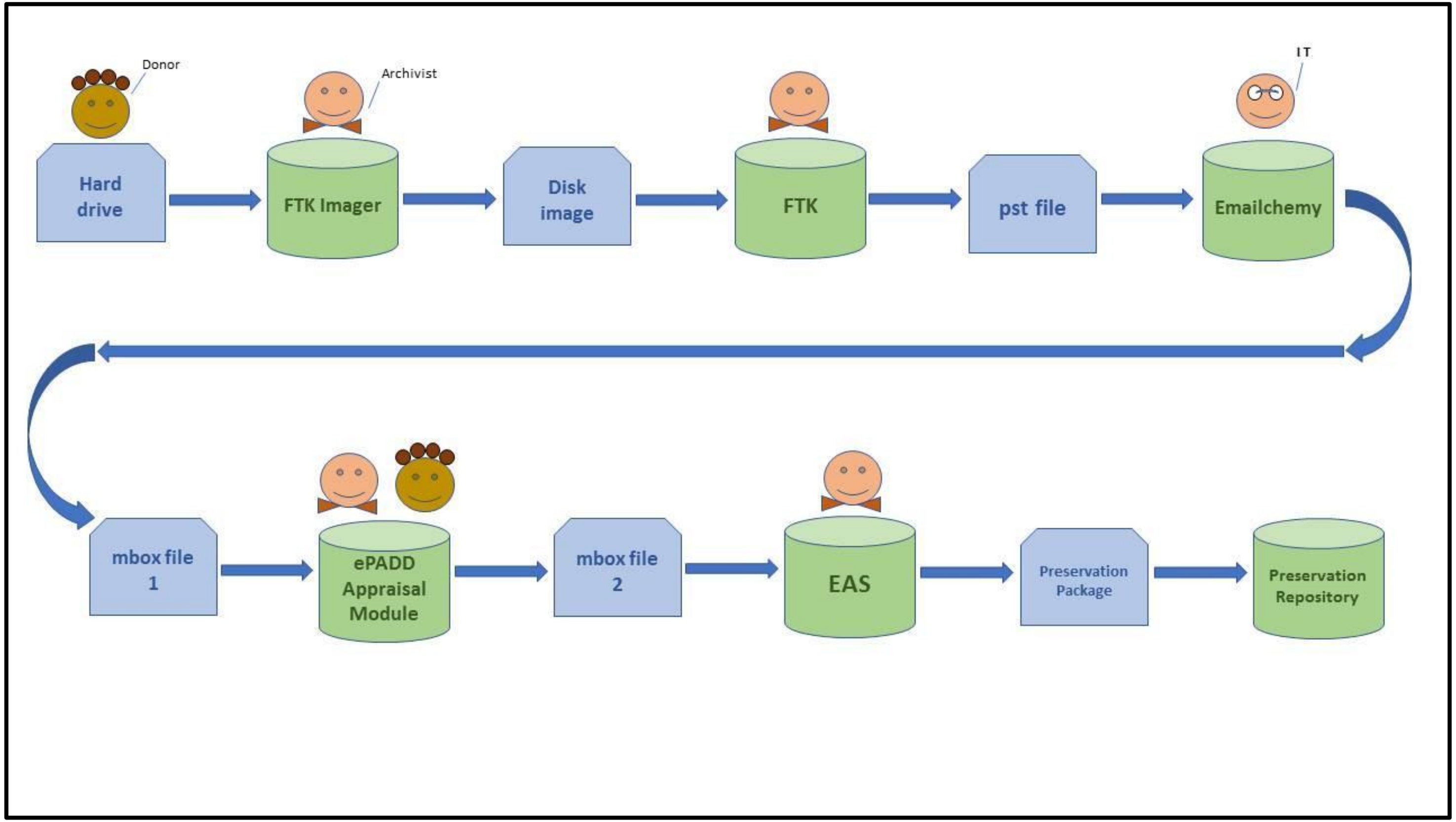
☐ Include mail in trash

☑ Search for mail location

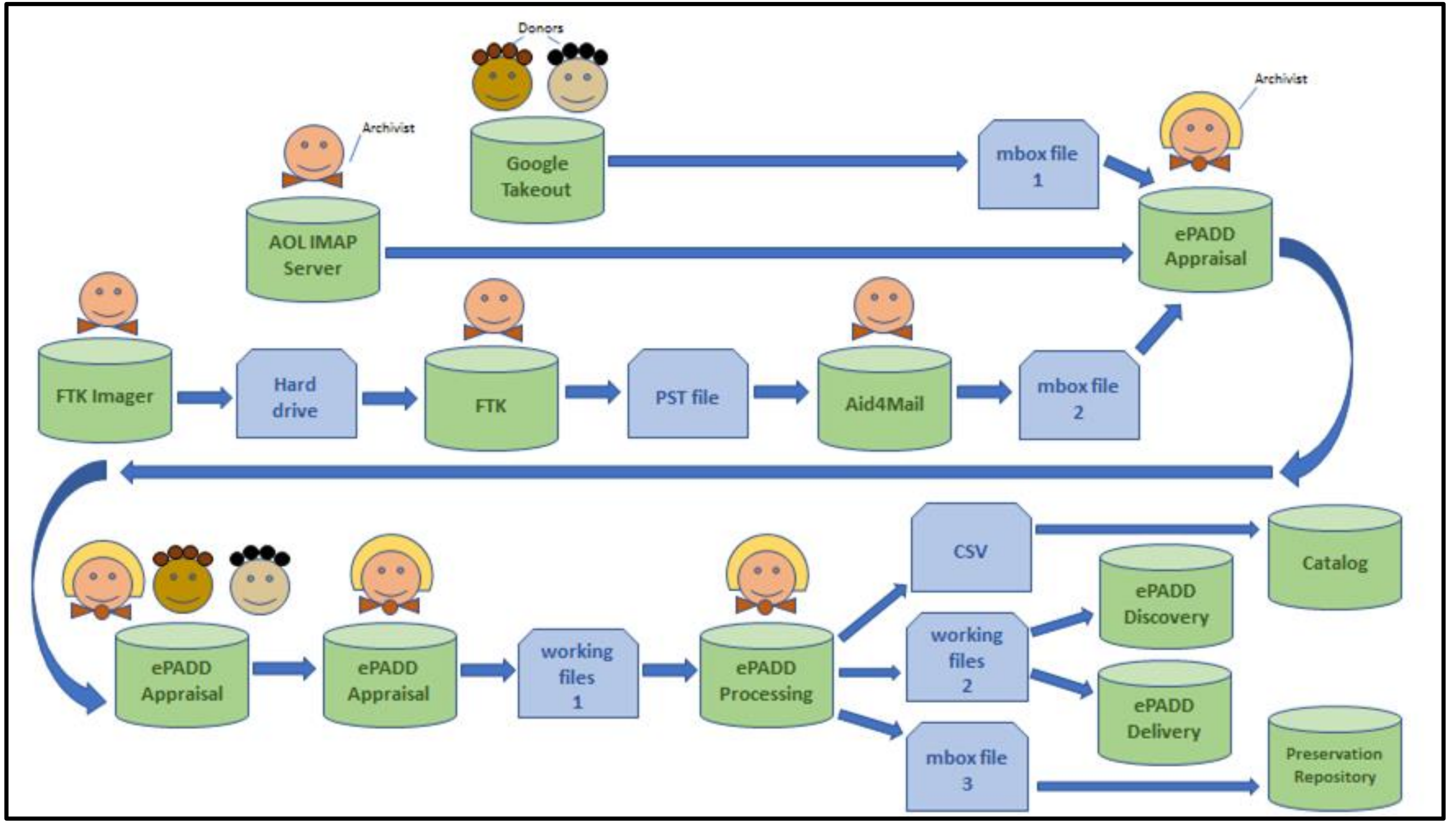
Profile name:

Aid4Mail Professional v3.3 -- Copyright © 2005-2014 Fookes Holding Ltd, Switzerland

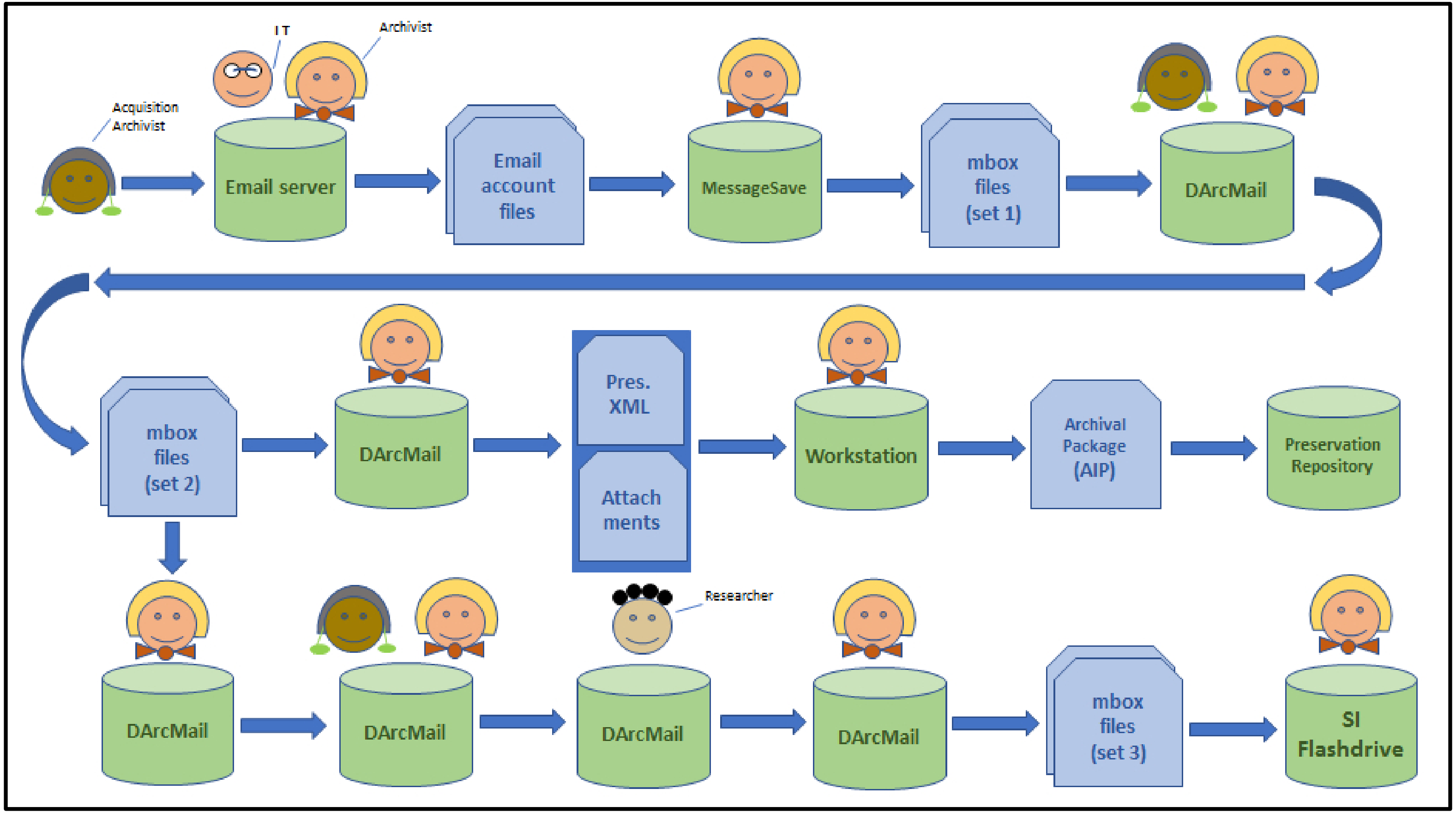
Workflow Scenario: Migration @ Harvard



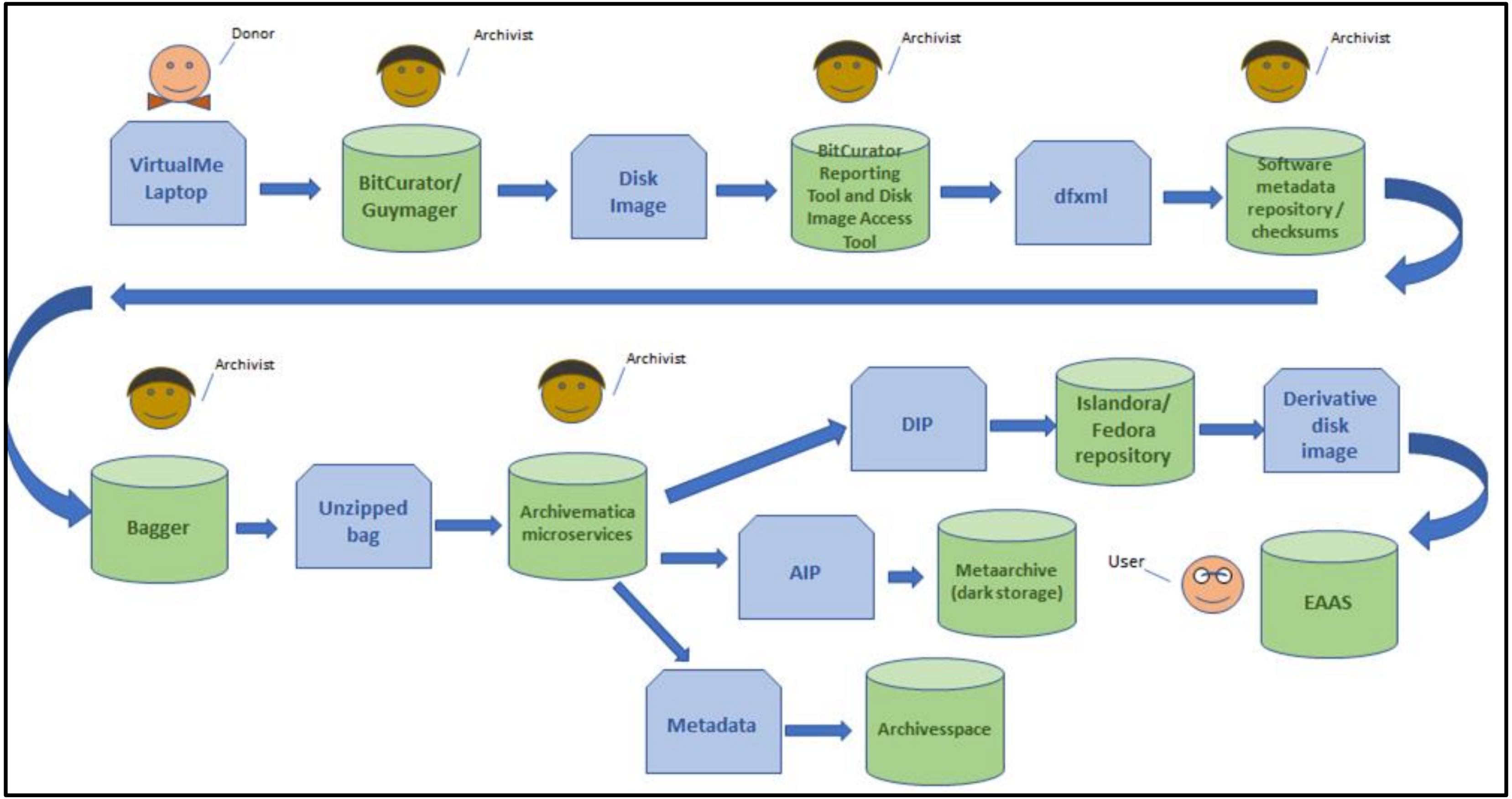
Migration @ Stanford with ePADD



Smithsonian XML Migration Scenario



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