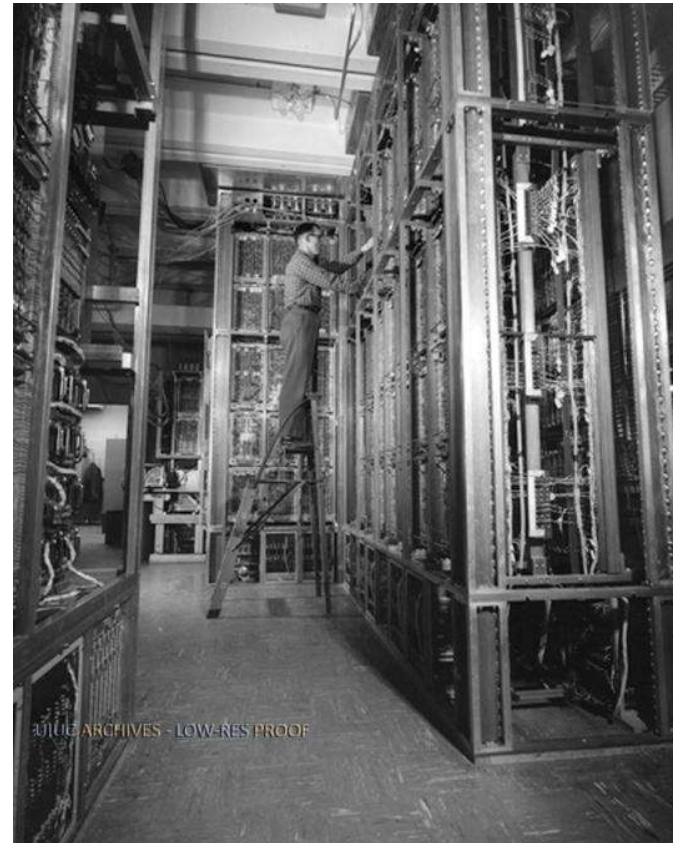


Working at Scale

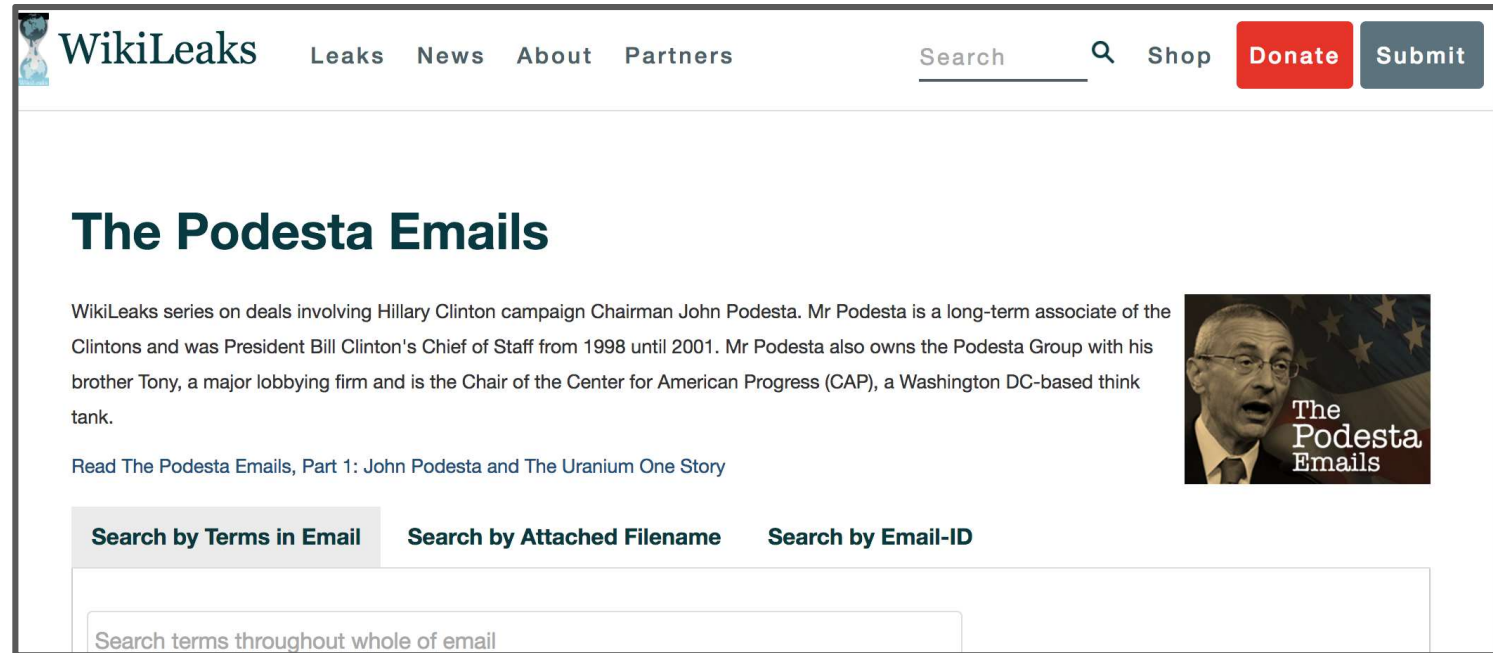
Use and number of
accounts

Size of account vis-à-vis
resources to preserve



**Iliac II and Harold Lopeman, ca 1958.
University of Illinois Archives, Computer
Science Photograph File.**

Security Issues



The screenshot shows the WikiLeaks website interface. At the top left is the WikiLeaks logo, followed by navigation links for 'Leaks', 'News', 'About', and 'Partners'. On the right side of the header, there is a search bar with a magnifying glass icon, a 'Shop' link, a red 'Donate' button, and a dark blue 'Submit' button. The main content area features the article title 'The Podesta Emails' in a large, bold, dark blue font. Below the title is a paragraph of introductory text: 'WikiLeaks series on deals involving Hillary Clinton campaign Chairman John Podesta. Mr Podesta is a long-term associate of the Clintons and was President Bill Clinton's Chief of Staff from 1998 until 2001. Mr Podesta also owns the Podesta Group with his brother Tony, a major lobbying firm and is the Chair of the Center for American Progress (CAP), a Washington DC-based think tank.' To the right of this text is a small image of John Podesta with the text 'The Podesta Emails' overlaid. Below the introductory text is a link: 'Read The Podesta Emails, Part 1: John Podesta and The Uranium One Story'. At the bottom of the article preview, there are three search filters: 'Search by Terms in Email' (highlighted in a light grey box), 'Search by Attached Filename', and 'Search by Email-ID'. Below these filters is a large search input field with the placeholder text 'Search terms throughout whole of email'.

Specific vulnerabilities include malicious content, spoofing, confidentiality, Third Party Privacy

Discovery and End User Access

In reading room?

Dedicated computer?

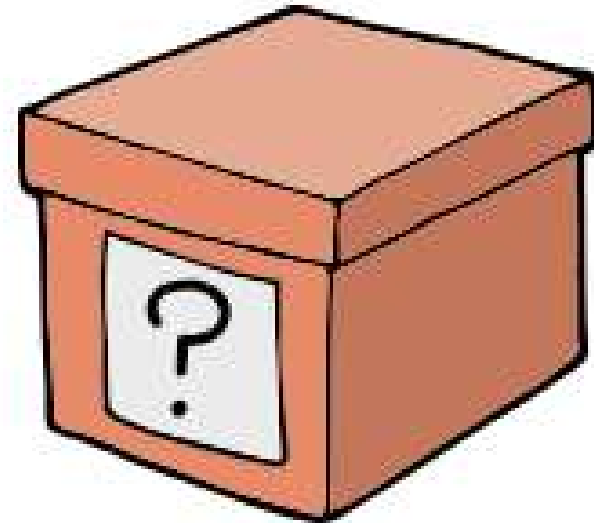
On web?

Redacted?

Aggregate Description?

Item level pathways?

Entire corpus?



Potential Solutions and Preservation Approaches



Preservation Approaches:

Begins with Account
and Format Analysis



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
 - Harvard faculty papers = minimum 50 years, can be up to 80 years
 - Princeton = 40 years as institutional record



Credit: Guy Levin, "An API First Development Approach," <https://dzone.com/articles/an-api-first-development-approach-1>

APIs and Interoperability

- APIs facilitate interoperability between systems
- Can accommodate different serialization and formats
- Can integrate systems without needing to know anything about underlying application language and functionality.
- Ability to leverage several layers or types of security and hide these from the user.
- Improved data integrity by letting the API ensure that it transferred data without corruption or loss
- Systems integration that uses APIs removes complexity for users and puts it in the application.

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?

Tools within Cultural Heritage Domain

- Key to interoperability, scalability, preservation and access
- Several are usable and maturing--with more work coming.



BitCurator

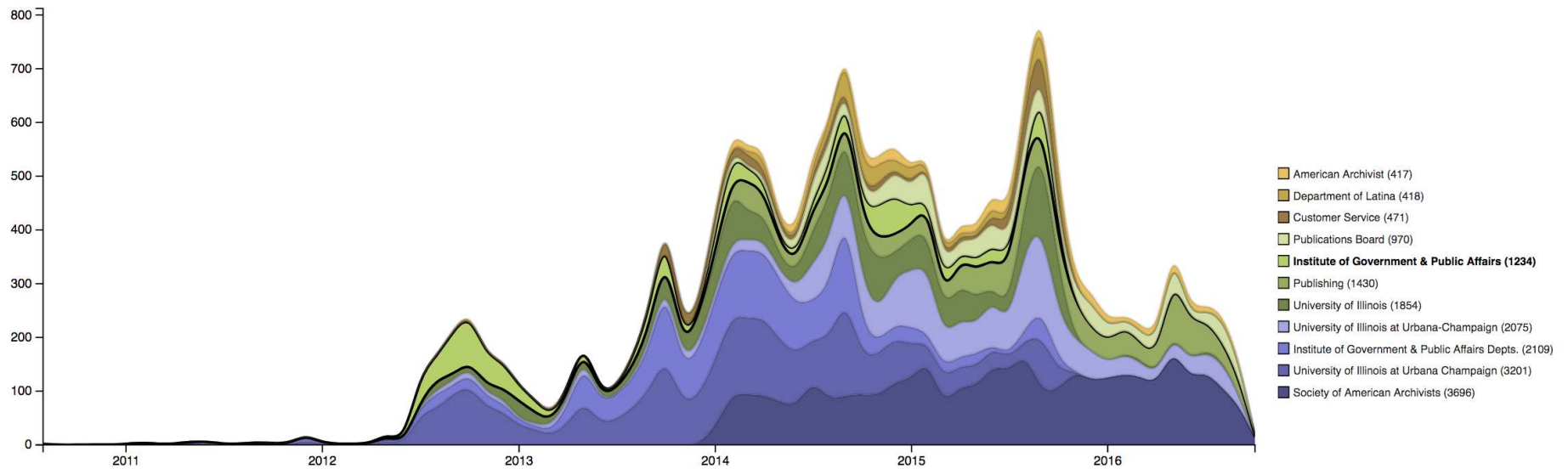
- Processing and Access of Disk Images
- Emulation Environments

EPADD: Process, Appraise, Discover, Deliver

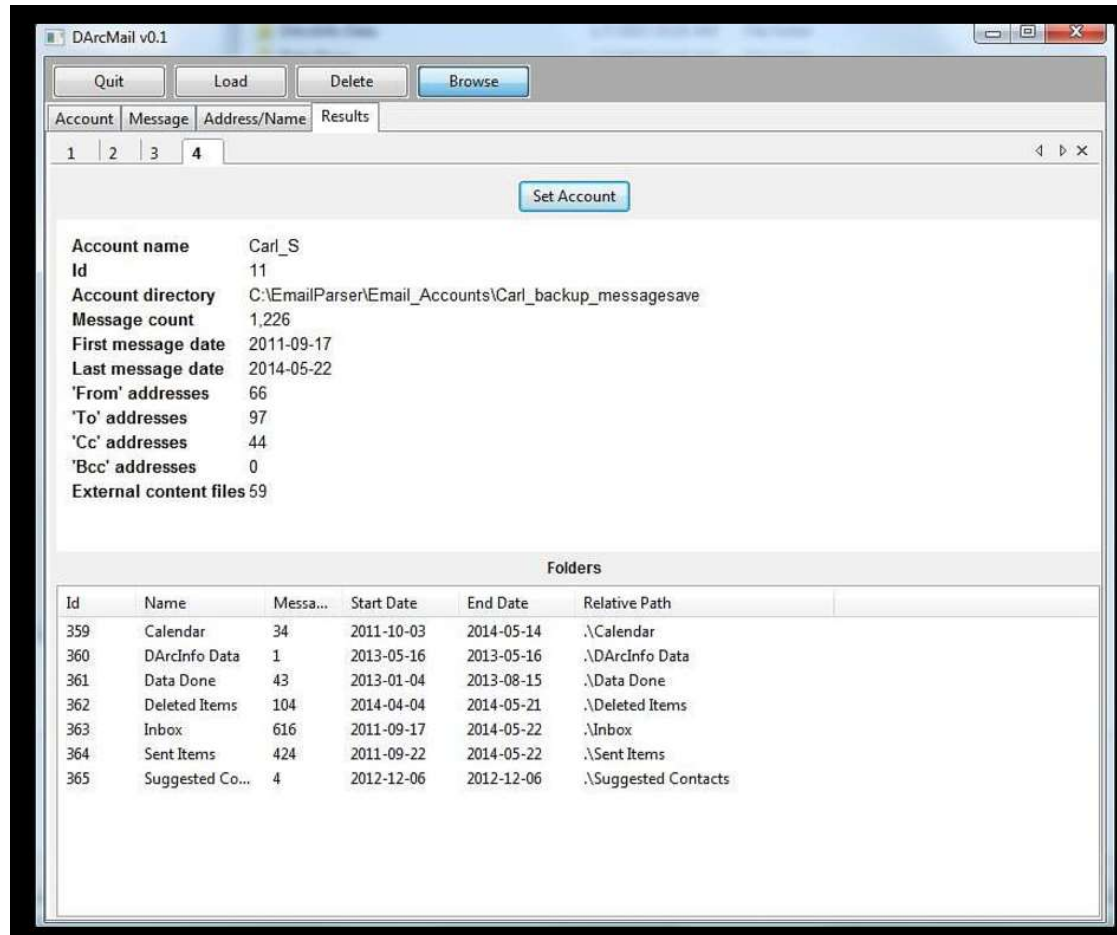


Jorge Chapa Email Series
Top entities graph (type: En_org)

[GO TO TABLE VIEW](#)



DArcMail



Harvard Electronic Archiving System - EAS

The screenshot shows the EAS web interface. At the top, there is a navigation bar with tabs for PACKETS, SEARCH, WORDSHACK, COLLECTIONS, ACCOUNTS, and SYSTEM INFO. The account is set to 'HUL TEST'. Below the navigation bar, there is a 'BRIEF DISPLAY' section with a search bar and filters. The search bar contains 'Search for' and 'Item type' is set to 'all'. The 'Results per page' is set to '10'. There are buttons for 'edit metadata', 'Delete', and 'Push to DRS'. Below the search bar is a table of search results. The table has columns for 'Select', 'EASi ID', 'Item type', 'From (display)', 'From (email)', 'To (display)', 'Date sent', and 'Subject'. The first few rows of the table are visible. At the bottom of the page, there is a pagination control showing '<< Previous 1 2 3 458 Next >>' and 'displaying 1 to 10 of 4578'.

Annotations on the screenshot:

- 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

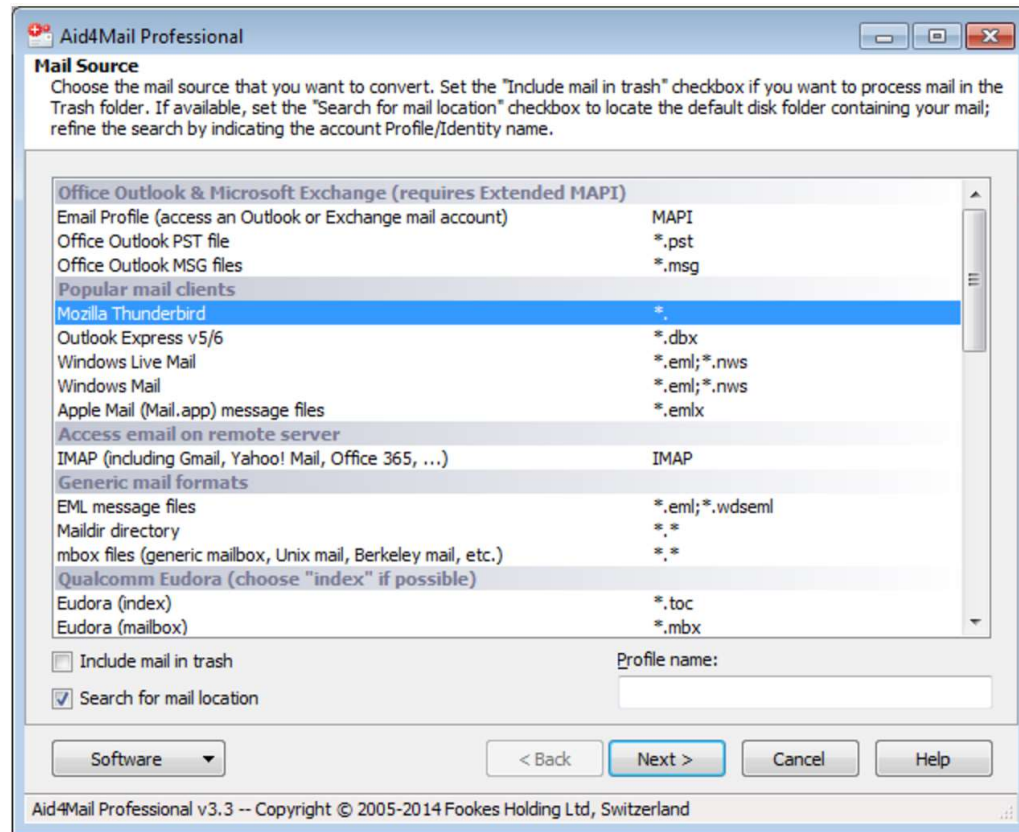
Select	EASi ID	Item type	From (display)	From (email)	To (display)	Date sent	Subject
<input type="checkbox"/>	55				Wendy Gogel	2009-07-31T22:19:12Z	Re: EMAP - minutes of June 30 meeting on de-dup
<input type="checkbox"/>	57		Skip Kendall	arvel_kendall@harvard.edu		2009-06-29T19:42:58Z	EMAP - scheduling more meetings
<input type="checkbox"/>	153		Skip Kendall	arvel_kendall@harvard.edu		2009-09-08T22:00:49Z	User requirements document
<input type="checkbox"/>	155		Wendy Gogel	wendy_gogel@harvard.edu	Susan von Salis	2009-08-06T18:21:53Z	Re: EMAP - minutes of June 30 meeting on de-dup
<input type="checkbox"/>	164					2009-06-19T19:48:00Z	EMAP - Meeting on Monday
<input type="checkbox"/>	169			wendy_gogel@harvard.edu		2009-06-25T15:43:50Z	Fwd: Re: Please update a email list
<input type="checkbox"/>	173		Wendy Gogel	wendy_gogel@harvard.edu	Andrea Goethals, Mar	2009-08-06T19:03:52Z	RE: EMAP - minutes of June 30 meeting on de-dup
<input type="checkbox"/>	175		Patti Fucci	patti_fucci@harvard.edu		2009-10-01T16:39:17Z	second meeting
<input type="checkbox"/>	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
<input type="checkbox"/>	182					2009-05-07T12:40:00Z	

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

Proprietary Tools

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

Also: Rest APIs



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

“Text as Data” Research Techniques

Named Entity Recognition (NER)

Machine learning

Topic modeling & dynamic topic modeling


Natural Language Processing (NLP)


Predictive Coding and Learning Classification


Continuous Active Learning (CAL)


KAINE 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.

 [Search the Collection](#)

 [Related Content](#)

 [Look Under the Hood](#)

 [What's the Buzz](#)

AutoTAR Technology-Assisted Review Platform

with Continuous Active Learning™ (CAL™)

Maura R. Grossman and Gordon V. Cormack



Search Jeb Bush Email:

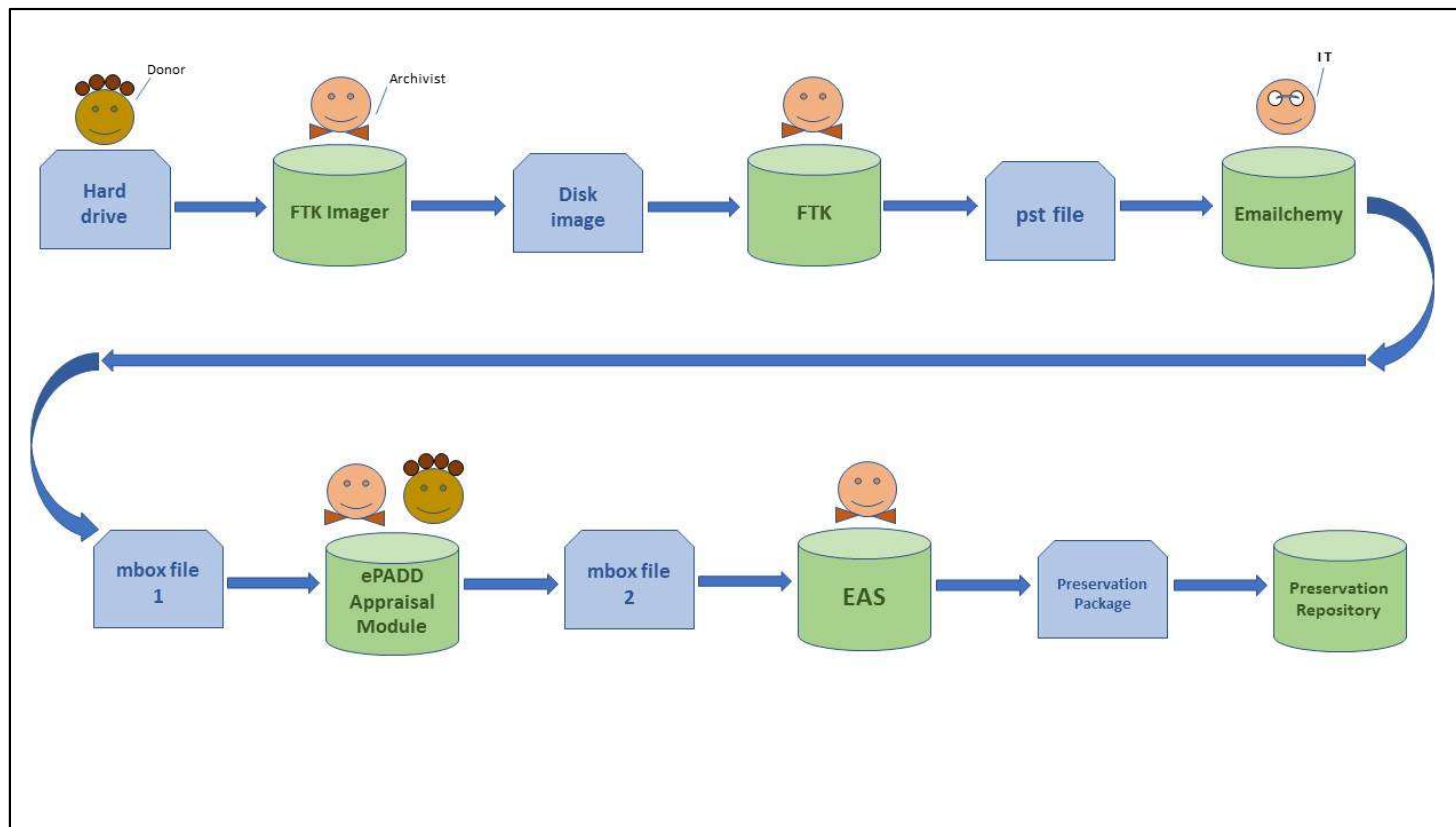
hotel

* CAL and Continuous Active Learning are Trademarks of [Maura R. Grossman](#) and [Gordon V. Cormack](#)

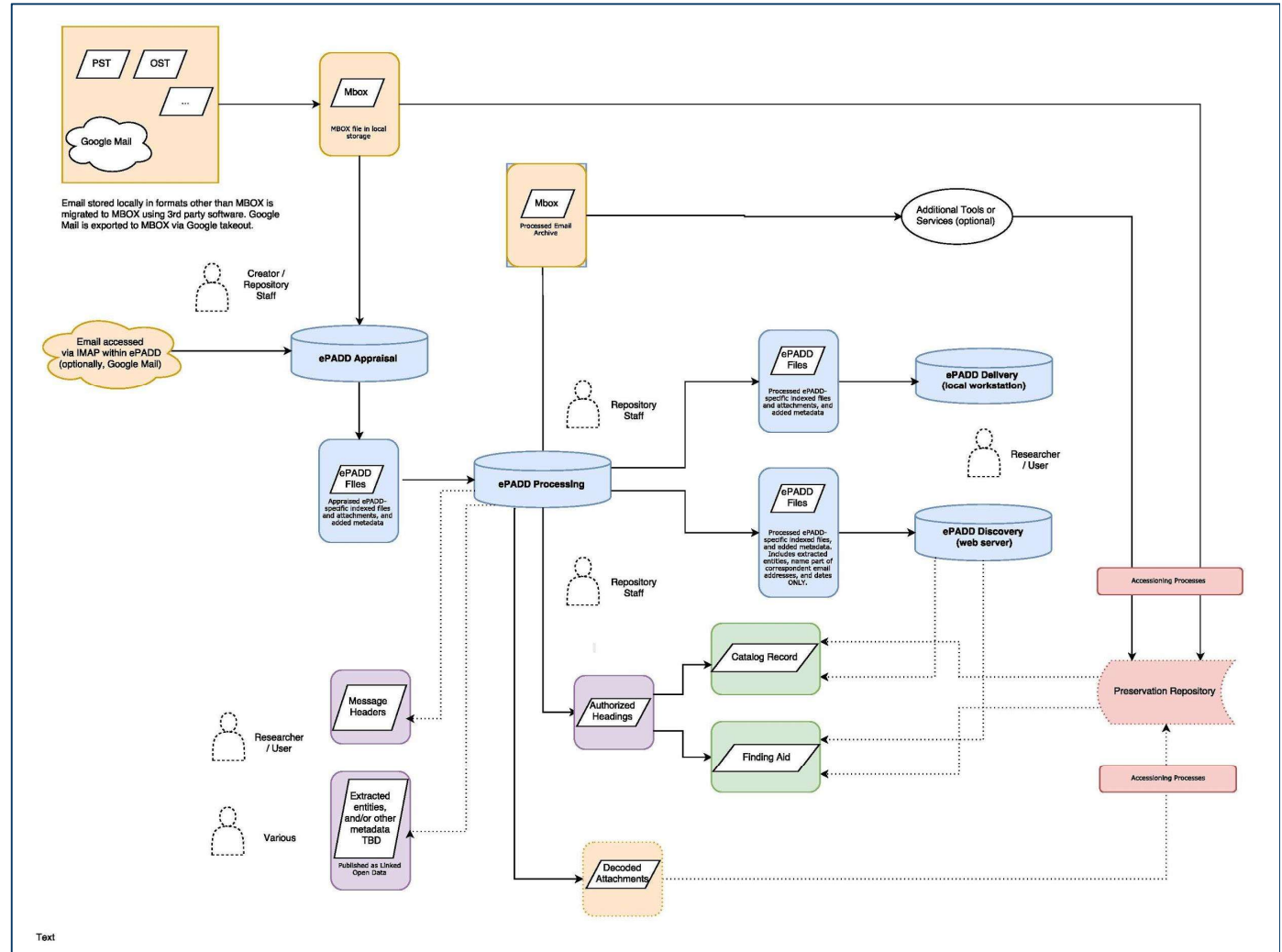
For more information on CAL, see our [Practical Law Journal Article](#)

Also available: [Search Tim Kaine Email](#)

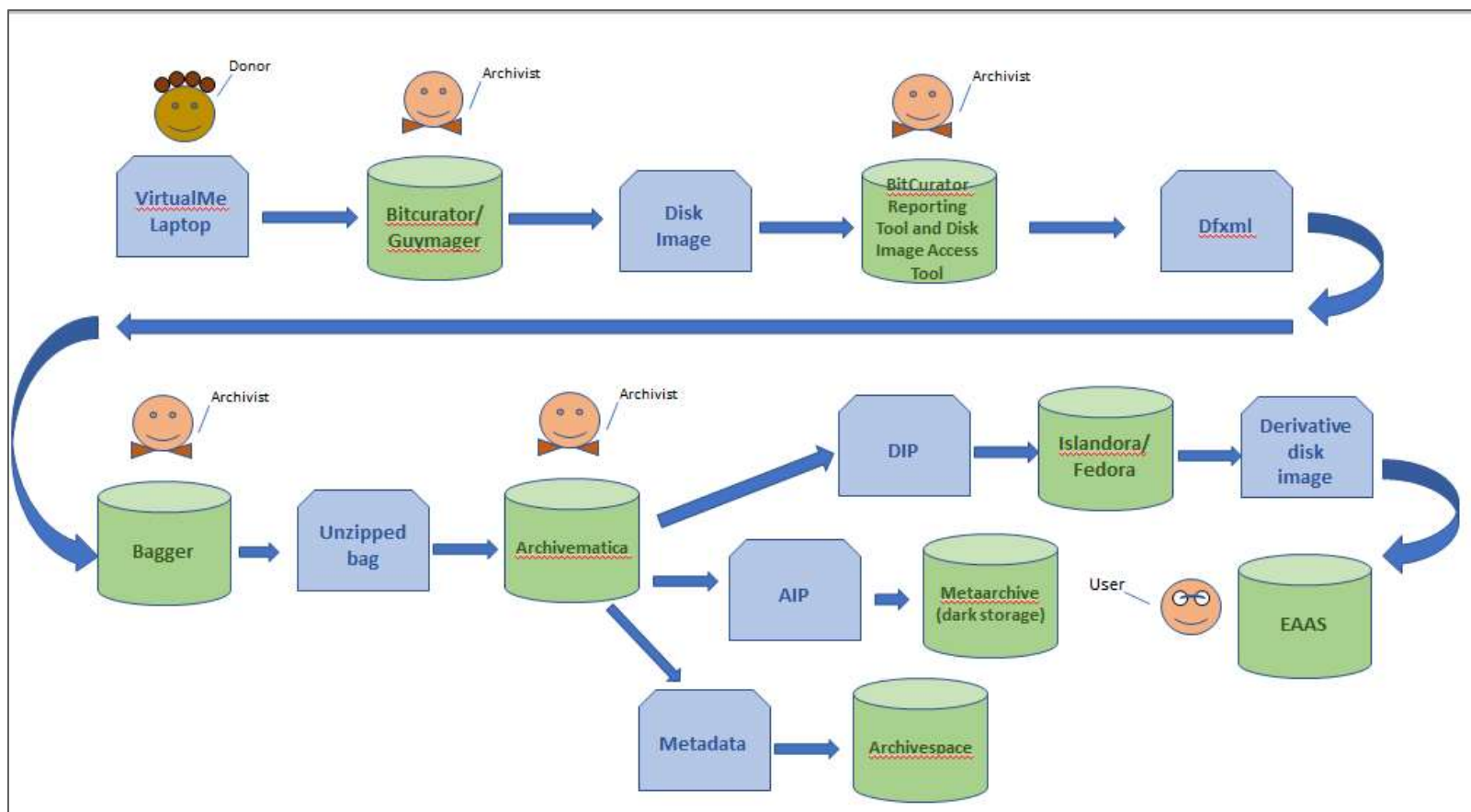
Workflow Scenario: Migration @ Harvard



Workflow Scenario: Migration @ Stanford with ePADD



WORKFLOW SCENARIO: ACCESSING AN ENTIRE DISK IMAGE - EMAIL & ATTACHMENTS - WITH EMULATION



What Gaps Do You
See?

Gaps and Recommendations: Format Standardization and Data Models

Improved documentation & standardization for MBOX, EML formats (& other formats)

Improve documentation for email data model, especially header fields

Improve tools for format identification & validation

Explore / test trace residue in tool changes, especially in supporting metadata and trail of actions taken by archivist or curator

Gaps and Recommendations for Authenticity

Improved tooling to test for completeness, non-alteration, and other aspects related to authenticity at different points of the preservation workflow

Opportunities to improved tooling (and methods generally) for header analysis to evaluate authenticity and integrity of messages based on just email data.

Gaps and Recommendations: Provenance & Chain of Custody

Ease to use use tools to facilitate harvesting of email from donors and depositors

Example: use of API's for continuing deposits

Process history data in machine actionable form - so need tools that can pull process history data as XML, then transform into PREMIS

Gaps and Recommendations: Collections at Scale

Mining automated signature blocks for name variation and association as well as providing context for matrixed relationships.

Use data corpus to improve authority work for named entities and correspondence

- Results testing of technologies and methods to automatically identify or redact PII/sensitive information
- Need to identify the staff skills needed to work with large scale digital collections.

Gaps and Recommendations: Metadata

What metadata elements (both technical and descriptive) should tools that archivists use support?

What metadata is adequately defined or documented?

How should it make that metadata available for integration into external systems?

How much of the transformation work that archivists undertake using email archiving tools be tracked and how should it be represented to the user.

Metadata for attachments

Gaps and Recommendations: Capture & Interoperability

Need more options for approaches for secure, reliable transfer of email collections, including capture via email APIs

Existence of well-documented API should be considered an essential feature for archival/curatorial tools

To facilitate use of APIs, devs should provide:

- Sandbox

- Standalone and documented libraries

- Support community development

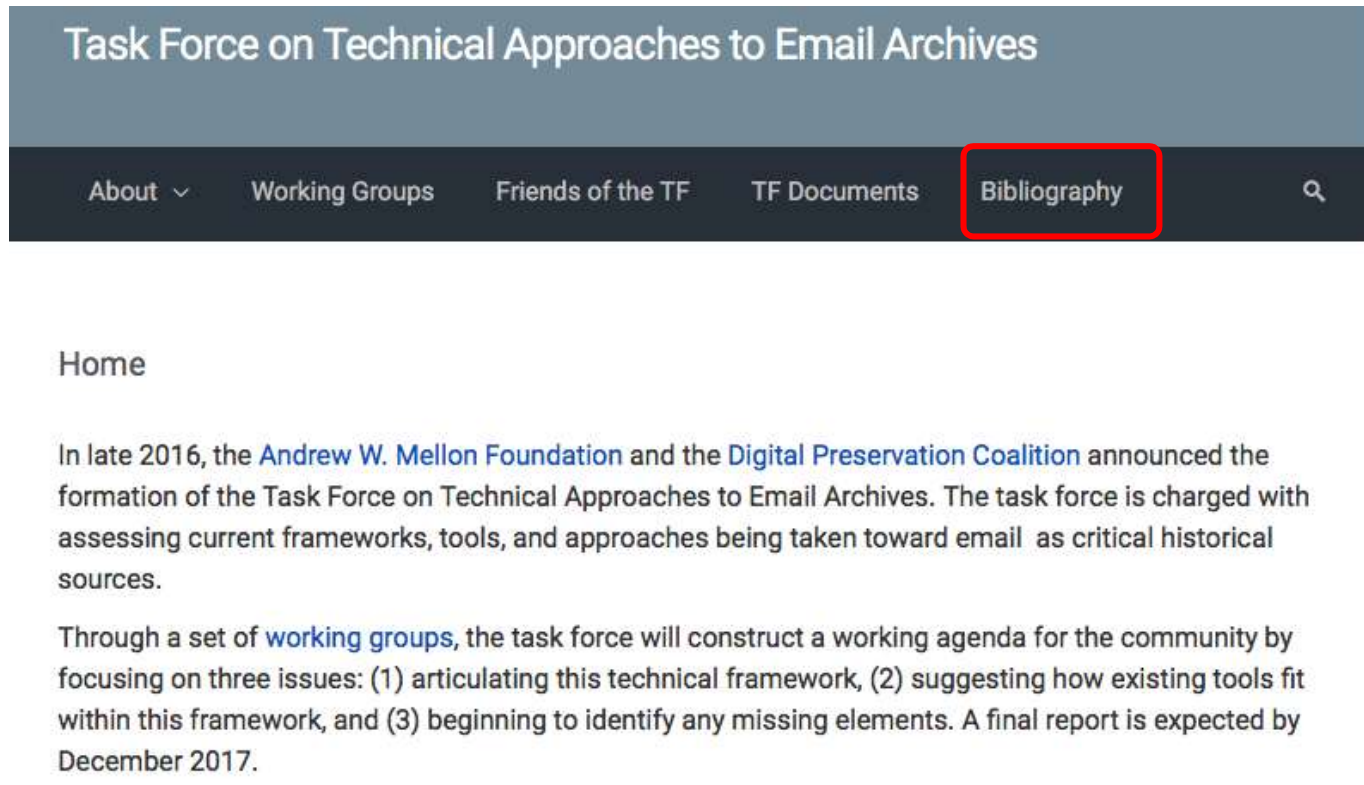
What's left to do in the report

Keep Drafting and editing

More workflow scenarios

Sustainability and Funder Roles

Deliverables: Website Resources



Task Force on Technical Approaches to Email Archives

About ▾ Working Groups Friends of the TF TF Documents **Bibliography** 🔍

Home

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.emailarchivestaskforce.org>

Bibliography

Task Force on Technical Approaches to Email Archives

About ▾ Working Groups Friends of the TF TF Documents **Bibliography** 🔍

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.alhhs.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.yip.to/immhf.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.

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

Date	Task	Status
October 2016	Launch meeting: Identify problem	Completed
November 2016 through April 2017	Working groups meet by phone/email to collaboratively draft text and formulate provisional findings and recommendations	Completed
April 2017	Stakeholder consultations including CNI, COSA (April 18), NAGARA, Museums and Web	Completed
June 30 2017	Draft report released for invited comment	Completed
July 6 2017	DPC Briefing Day to get UK input on early drafts	Right now!
July 2017 through August 2017	Working groups incorporate feedback and continue drafting; Co-chairs and assistant edit draft consultation report	In process
August through October 2017	Open feedback solicited and incorporated into the report	To come
September 2017	Executive Committee Meeting	To come
December 2017	Full task Force meeting in New York to conclude work and approve report	To come
January 2018	Task Force Report Published, DPC Briefing Day	To come

Questions, Feedback, Discussion

Chris Prom: prom@illinois.edu

Kate Murray: kmur@loc.gov

<http://www.emailarchivestaskforce.org>