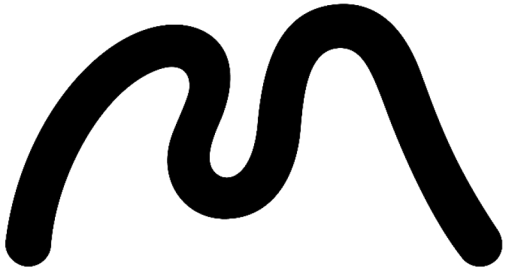
PDF/Mail – Format Overview, Tool Architecture, and Future Directions

Chris Prom and Tom Habing
University of Illinois at Urbana-Champaign

Contributions from
Peter Wyatt, PDF Association
Eden Irwin and Ruby Martinez, University of Illinois at Urbana-Champaign

Email Archiving Symposium
June 15, 2023

INSTITUTE *of*
Museum and Library
SERVICES



Mellon
Foundation

Phase One

Support from
Andrew W. Mellon Foundation

Chris Prom, University of Illinois (PI)

Kevin De Vorsey - National Archives and Records Administration

Kate Murray - Library of Congress

Lynda Schmitz Fuhrig - Smithsonian Archives

Steve Levenson - ISO TC 171 SC2 WG5 Convenor for PDF/A

Stephen Abrams - Harvard University Libraries

Academic/industry working group

Tricia Patterson - Harvard University Libraries

Cal Lee, UNC School of Information and Library Science

Camille Tyndall Watson - State Archives of NC

Jamie Patrick-Burns - State Archives of NC

Duff Johnson - PDF Association

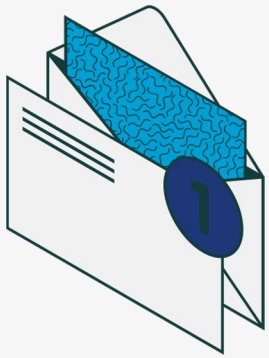
Matthew Hardy - Adobe Systems Inc.

Dietrich von Seggern, Callas software, GmbH

Joel Simpson - Artefactual Systems

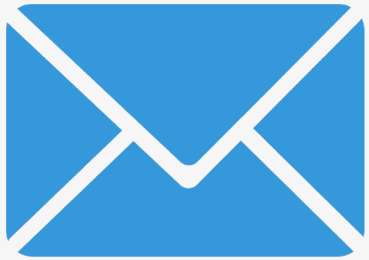
ILLINOIS
University Library

Why Preserve Email with PDF?



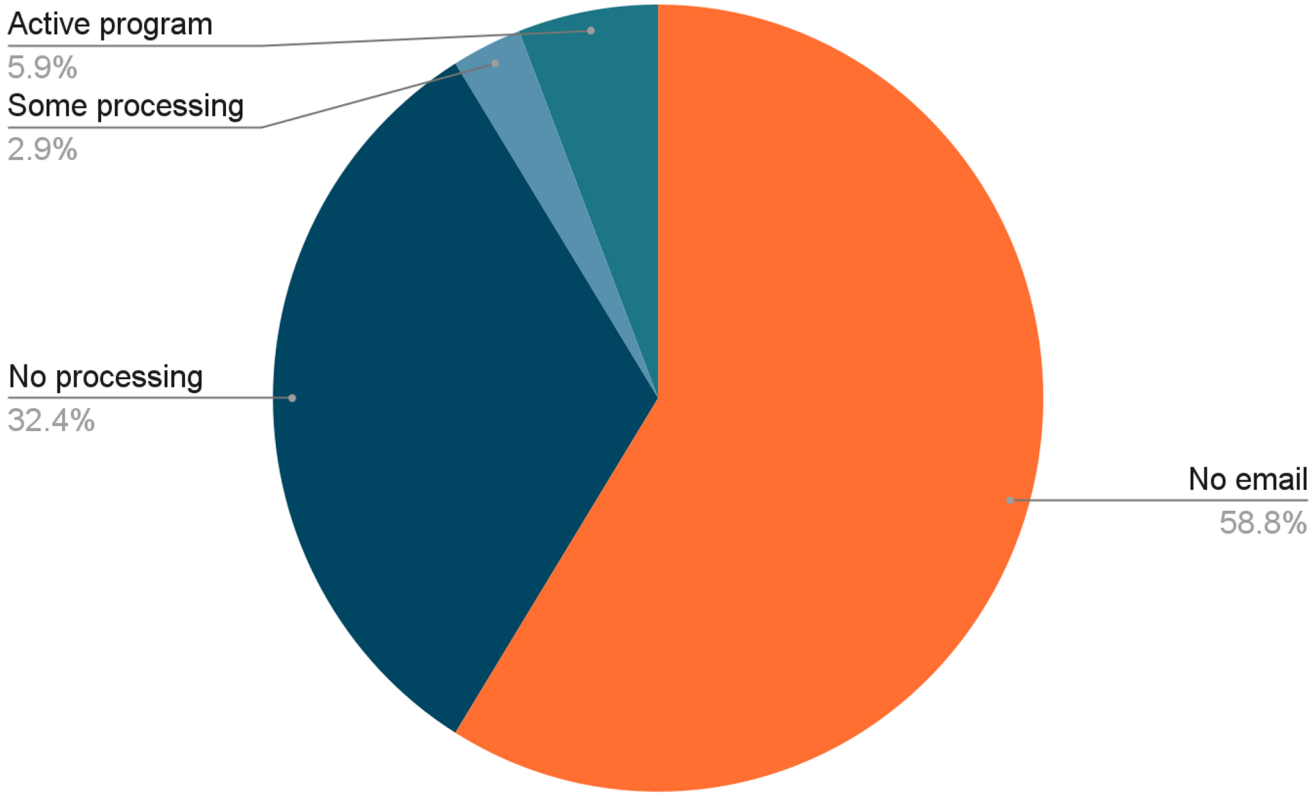
Email is one of the most ubiquitous forms of communication

Email is historical evidence

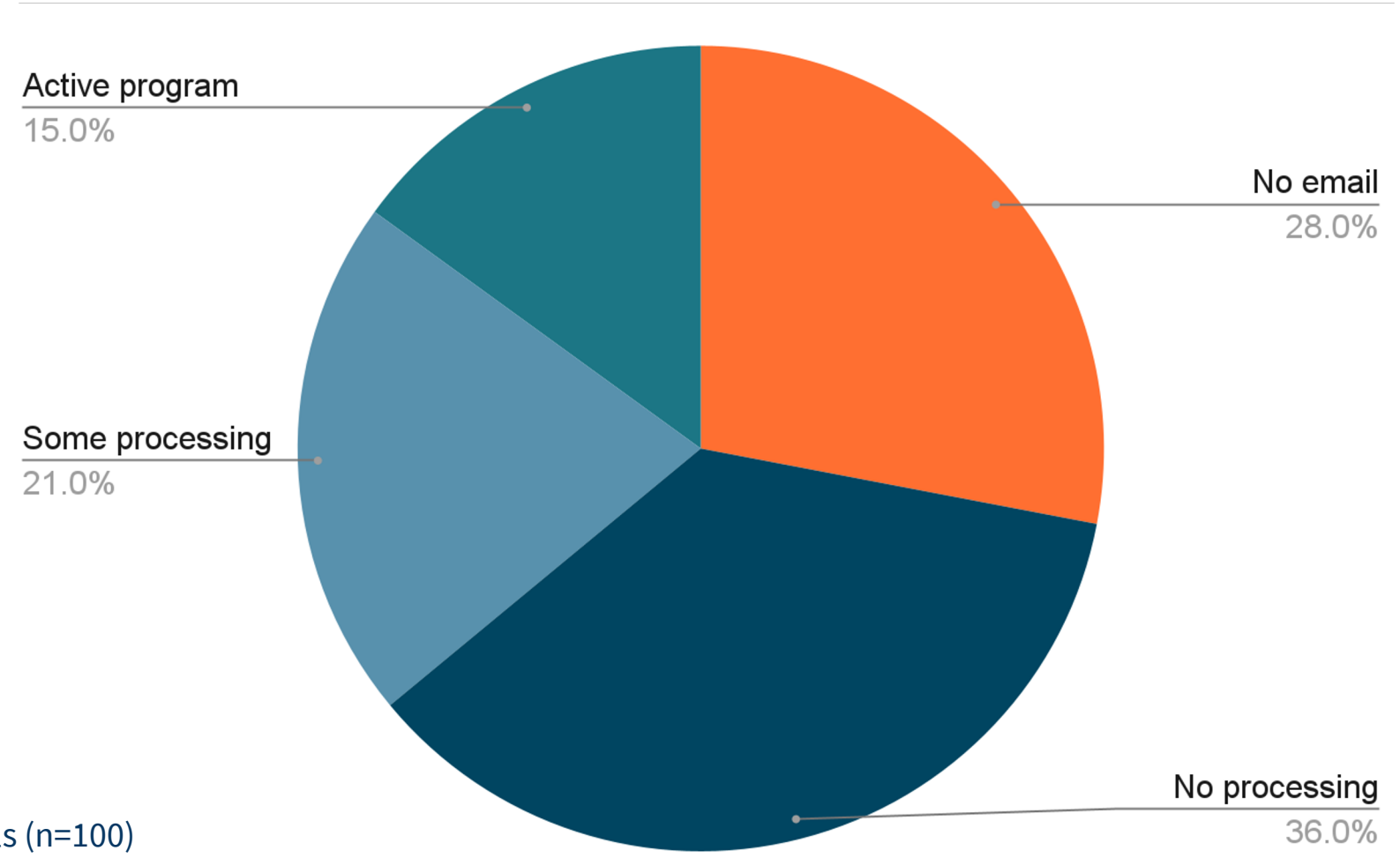


Email holds an incredible amount of information

Current State of Email Archiving (State of Illinois Sample)



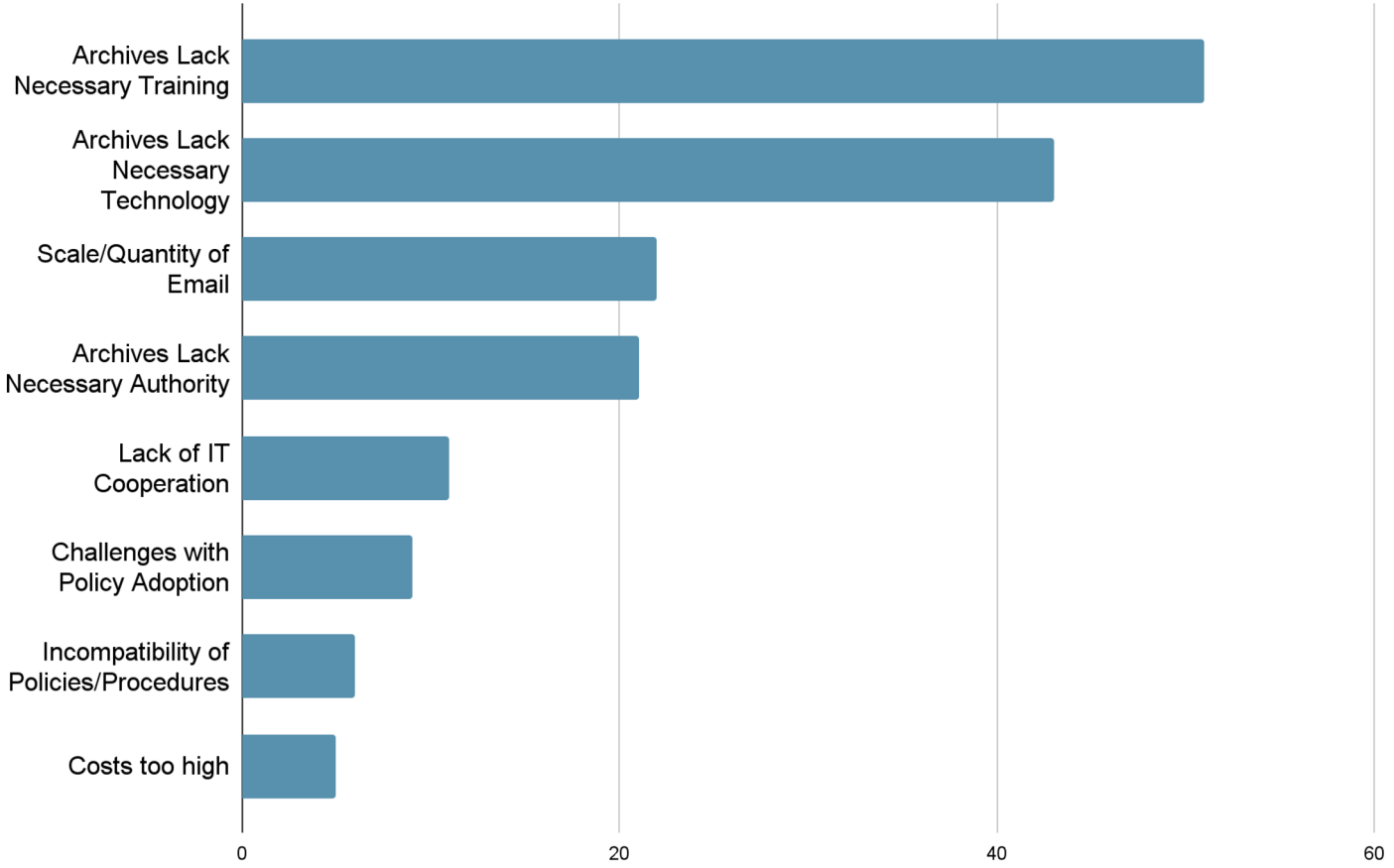
Survey (n=68)



Live Polls (n=100)

ILLINOIS
University Library

Why are archives not currently preserving email?



ILLINOIS University Library

Demand for the PDF alternative

Software/Formats used for Email Archives, Survey of Archives in Illinois (n=68)	
NONE	43
PDF	8
EMAIL APPLICATIONS	6
NOT SURE	5
PRESERVICA	3
CURRENTLY CHOOSING	2
OPEN SOURCE TOOLS	1

ILLINOIS
University Library

Why PDF/Mail?

Complementary to preservation features of PDF/A

Standardizes metadata and evidential value

PDF is already used within archive community

ILLINOIS
University Library

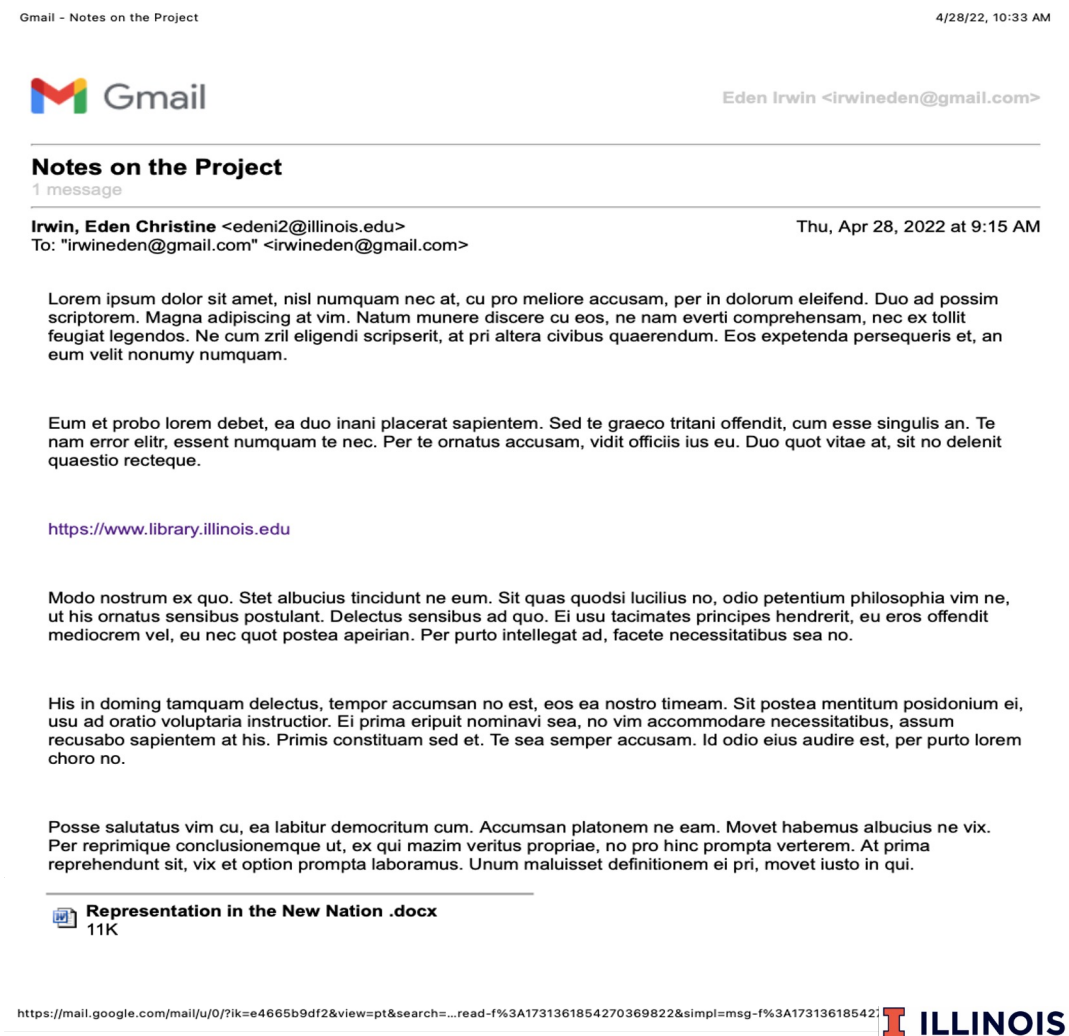
Current PDF Functionality

Visible Message
The header is visible, as is the body of the message.

Internal Information
The only metadata that remains is about the created PDF, not the email

Attachments and Bulk Conversion
Neither can be expected

Links
Links remain active but content not included



Gmail - Notes on the Project 4/28/22, 10:33 AM

Gmail Eden Irwin <irwineden@gmail.com>

Notes on the Project
1 message

Irwin, Eden Christine <edeni2@illinois.edu> Thu, Apr 28, 2022 at 9:15 AM
To: "irwineden@gmail.com" <irwineden@gmail.com>

Lorem ipsum dolor sit amet, nisl numquam nec at, cu pro meliore accusam, per in dolorum eleifend. Duo ad possim scriptorem. Magna adipiscing at vim. Natum munere discere cu eos, ne nam everti comprehensam, nec ex tollit feugiat legendos. Ne cum zril eligendi scripserit, at pri altera civibus quaerendum. Eos expetenda persequeris et, an eum velit nonumy numquam.

Eum et probo lorem debet, ea duo inani placerat sapientem. Sed te graeco tritani offendit, cum esse singulis an. Te nam error elitr, essent numquam te nec. Per te ornatus accusam, vidit officiis ius eu. Duo quot vitae at, sit no delenit quaestio recteque.

https://www.library.illinois.edu

Modo nostrum ex quo. Stet albucius tincidunt ne eum. Sit quas quodsi lucilius no, odio petentium philosophia vim ne, ut his ornatus sensibus postulant. Delectus sensibus ad quo. Ei usu tacimates principes hendrerit, eu eros offendit mediocrem vel, eu nec quot postea apeirian. Per purto intellegat ad, facete necessitatibus sea no.

His in doming tamquam delectus, tempor accumsan no est, eos ea nostro timeam. Sit postea mentitum posidonium ei, usu ad oratio voluptaria instructior. Ei prima eripuit nominavi sea, no vim accommodare necessitatibus, assum recusabo sapientem at his. Primis constituam sed et. Te sea semper accusam. Id odio eius audire est, per purto lorem choro no.

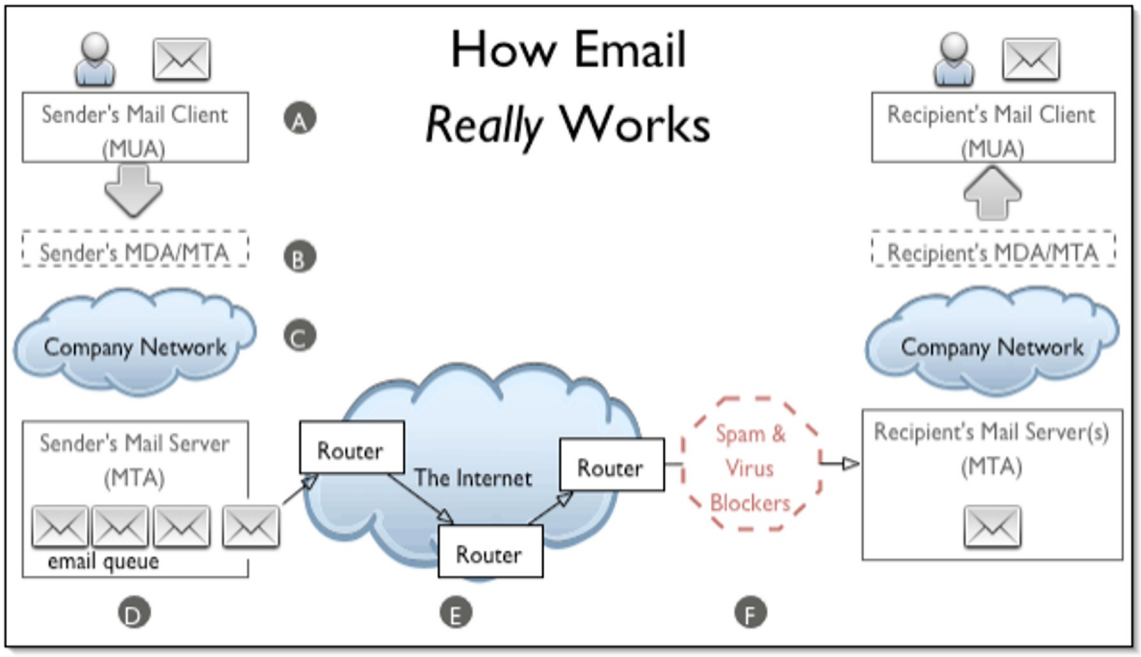
Posse salutatus vim cu, ea labitur democritum cum. Accumsan platonem ne eam. Movet habemus albucius ne vix. Per reprimique conclusionemque ut, ex qui mazim veritus propriae, no pro hinc prompta verterem. At prima reprehendunt sit, vix et option prompta laboramus. Unum maluisset definitionem ei pri, movet iusto in qui.

Representation in the New Nation .docx
11K

https://mail.google.com/mail/u/0/?ik=e4665b9df2&view=pt&search=...read-f%3A1731361854270369822&simpl=msg-f%3A173136185427

ILLINOIS
University Library

Email Fixity for Archives in PDF



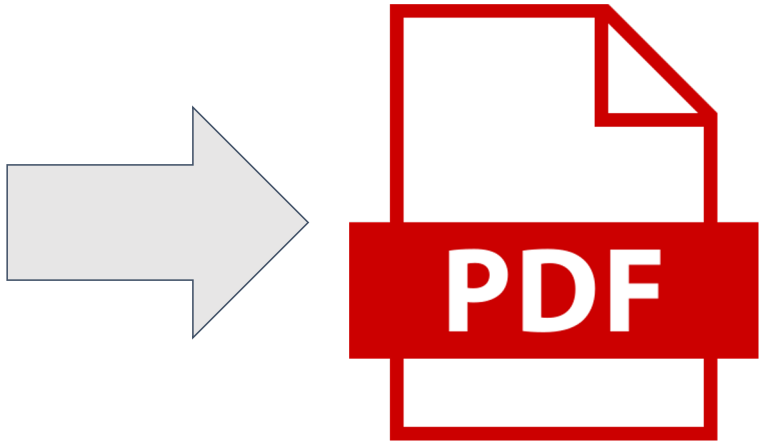


Image via: https://www.oasis-open.org/khelp/kmlm/user_help/html/how_email_works.html

Phase One Report

- Articulates rationale for EA-PDF (PDF/Mail)
- Define conceptual requirements for (PDF/Mail) container: a PDF file containing email data in defined structures and having several core archival attributes.
- Describe functional requirements for PDF/Mail-specific viewers.

ILLINOIS
University Library

Phase Two

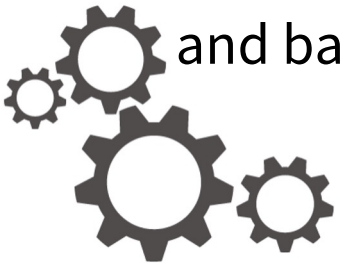
INSTITUTE of Museum and Library SERVICES

ILLINOIS University Library

PDF association

Goals of Phase Two

- Collaboration: Academic/industry partnership
- Specification: A detailed technical description for the EA-PDF (email archives in PDF) file format
- Tool-building: A proof-of-concept, open-source EA-PDF writer and baseline expectations for PDF/M aware viewer

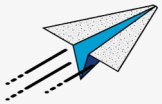


ILLINOIS
University Library

PDF Association – pdfa.org

Current Work

Liaison Working Group Develops Draft Spec for Comment



UIUC Leads Tool Development

Proof of concept tool that Tom will describe

Project Fosters Collaboration, Testing

Tool Testing

Community Building

Key features of PDF/mail

- Core Representation and Fields
- Embedded files (compressed)
  - The “raw” email file (EML, MBOX, PST, …)
  - Email attachments
  - HTML assets
- XMP-based metadata
- Navigational aids
  - Outlines, file attachment annotations, tagged PDF?

PDF/mail creation software

Legacy viewer vs. PDF/mail viewer

EA-PDF XMP Metadata

EA-PDF Profile	Description
PDF/mail-1s (single)	A single email preserved as a single EA-PDF file → metadata describes that single email `Email{ Subject, To, Cc, Bcc, … }`
PDF/mail-1m (multiple)	A single EA-PDF file containing many emails (non-hierarchical) → metadata describes multiple emails, email by email `Email1{Subject, To, Cc, …}, Email2{Subject, To, CC, …}, …`
PDF/mail-1c (container)	A “structured container” for a multiple EA-PDF files forming a hierarchy, each of which is a PDF/mail-1s, PDF/mail-1m or PDF/mail-1c → metadata describes the container, not the emails

XMP for EA-PDF

Mandatory - at document (file) level

- Exactly like PDF/A → *since we want PDF/A compatibility*
- XMP is "best practice" for modern PDF
  - Document Catalog Metadata for the document (file)
  - A lot of "generic" legacy software does *not* provide access to XMP metadata

EA-PDF XMP includes core email header files (Subject, To, From, CC, BCC, …)

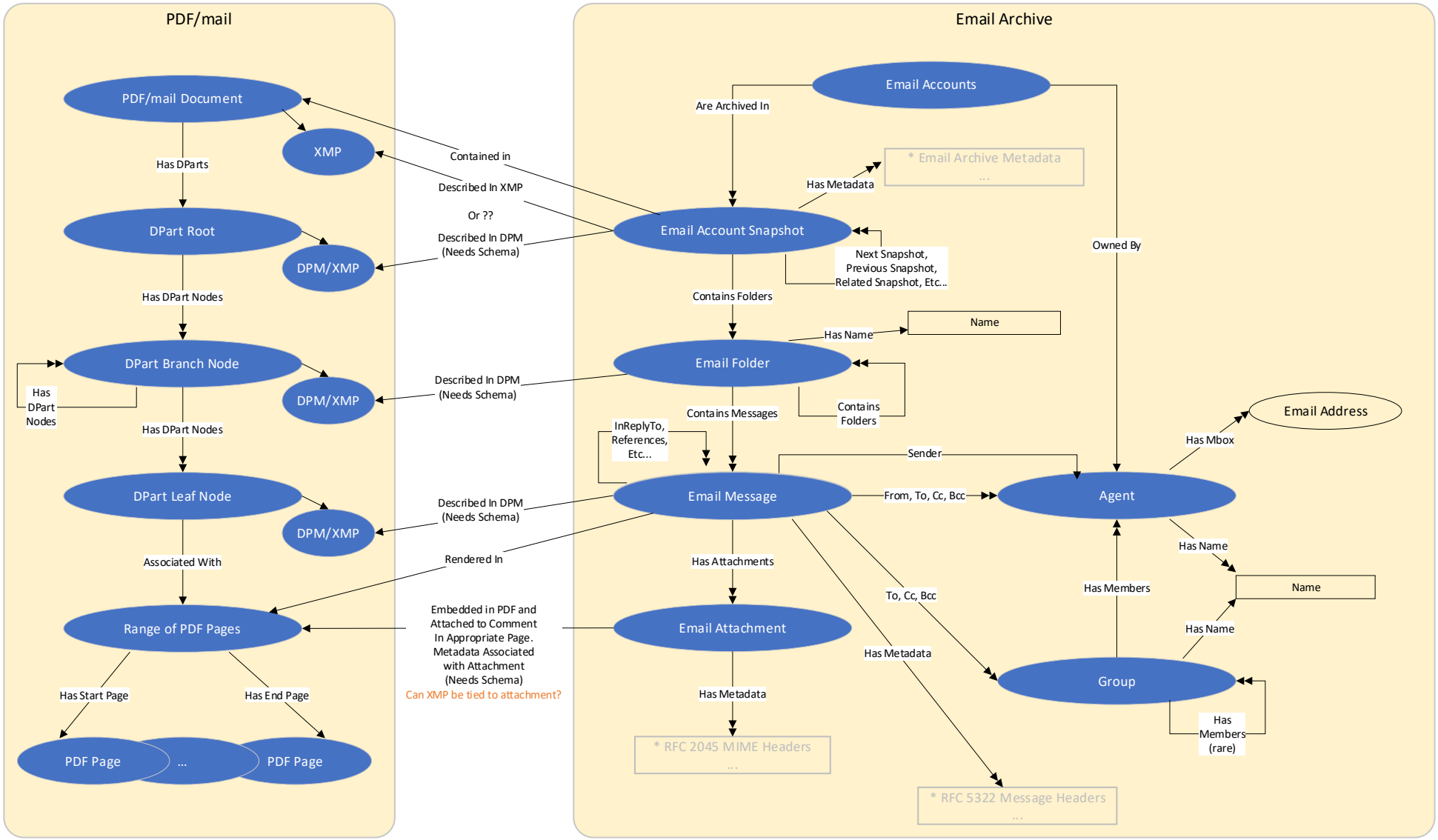
- For better discovery, search, etc. by *existing* document management systems
  - Not specifically EA-PDF aware, but assumed to be XMP aware (e.g. for PDF/A and other formats)

EA-PDF Technology

- **EA-PDF**: encapsulating concepts for preservation of email
- **PDF/mail**: PDF Association *subset* specification for preservation of email
  - *PDF/mail is under development!*
- **PDF/M**: potential *future* ISO equivalent to **PDF/mail**
- Based on **PDF/A-3 (PDF 1.7)** and **PDF/A-4f (PDF 2.0)**
- Profiles (conformance levels)
  - PDF/mail-1s (single email = EML)
  - PDF/mail-1m (multiple emails = MBOX)
  - PDF/mail-1c (container of many PDF/mail files = PST)

Technical Overview/Points

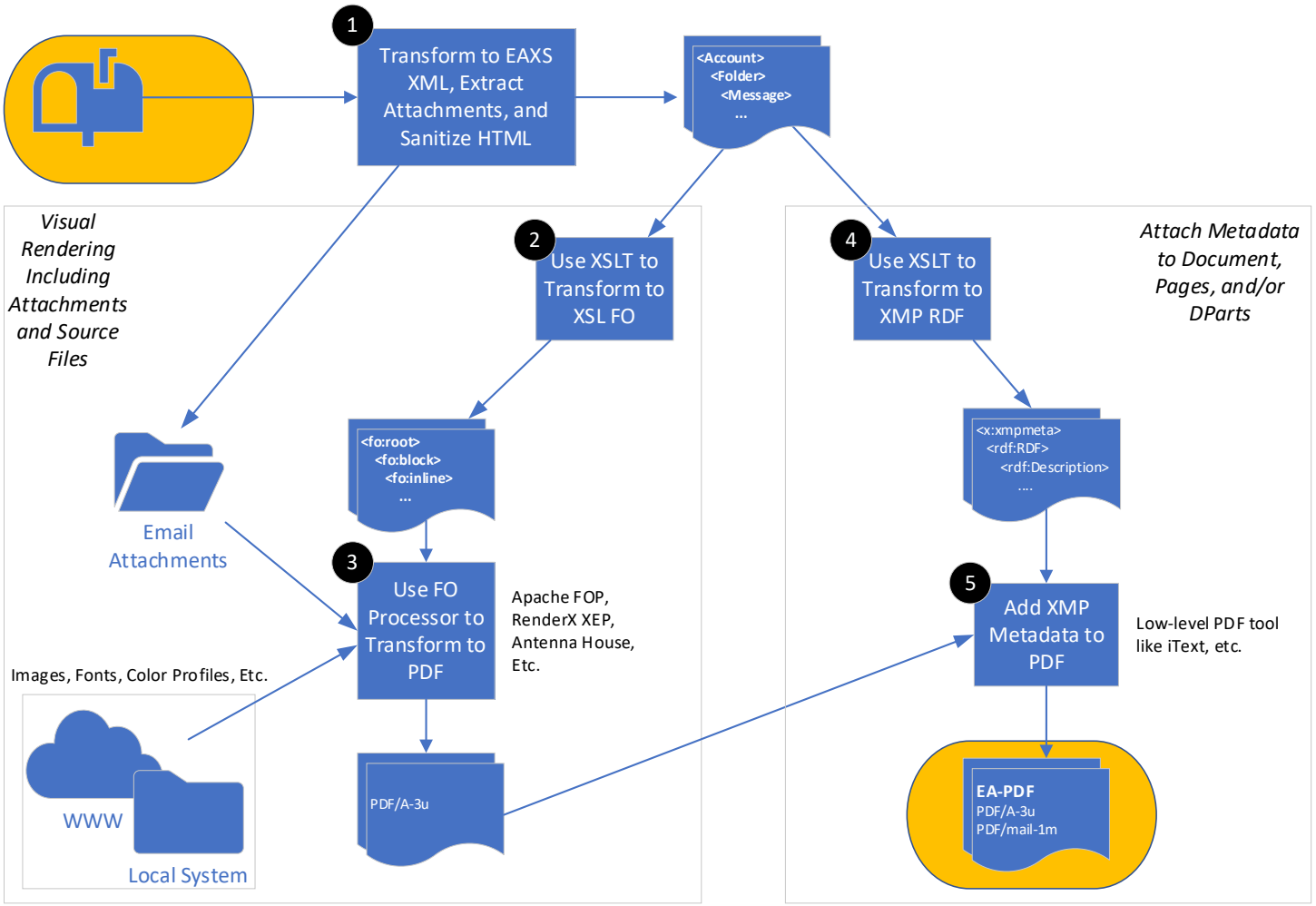
- PDFs must be archival quality, PDF/A-3 (1.7) or PDF/A-4 (2.0)
- One or more emails per PDF file
  - PDF Portfolios may be used
- Combination of human-readable and machine-readable content
  - Core email headers plus text content (plain and/or html) are human readable
    - May also include TOCs, summary data, conversion warnings, etc.
    - Available in legacy PDF viewers
  - Metadata at the corpus, folder, and message-level is machine-readable
    - Current proposal is to use the PDF DParts Specification (next slide)
    - Custom metadata format for email headers plus DACS-compatible fields (under development)
    - May not all be accessible in legacy PDF viewers
- All email attachments are embedded in the PDF as PDF attachments
- Also, original source emails will be embedded as PDF attachments
- Show and describe an example PDF


PDF/mail
PDF/mail Document
XMP
Has DParts
DPart Root
DPM/XMP
Has DPart Nodes
DPart Branch Node
Has DPart Nodes
DPart Leaf Node
DPM/XMP
Associated With
Range of PDF Pages
Has Start Page
Has End Page
PDF Page
...
PDF Page
Contained in
Described In XMP
Or ??
Described In DPM (Needs Schema)
Described In DPM (Needs Schema)
Described In DPM (Needs Schema)
Rendered In
Embedded in PDF and Attached to Comment In Appropriate Page. Metadata Associated with Attachment (Needs Schema)
Can XMP be tied to attachment?
Email Archive
Email Accounts
Are Archived In
Owned By
Has Metadata
* Email Archive Metadata
...
Email Account Snapshot
Next Snapshot, Previous Snapshot, Related Snapshot, Etc...
Contains Folders
Has Name
Name
Email Folder
Contains Folders
Contains Messages
InReplyTo, References, Etc...
Sender
Email Message
From, To, Cc, Bcc
Agent
Has Mbox
Email Address
Has Attachments
Email Attachment
Has Metadata
* RFC 2045 MIME Headers
...
To, Cc, Bcc
Has Metadata
* RFC 5322 Message Headers
...
Has Members
Has Name
Name
Has Name
Group
Has Members (rare)

Proof of Concept Implementation

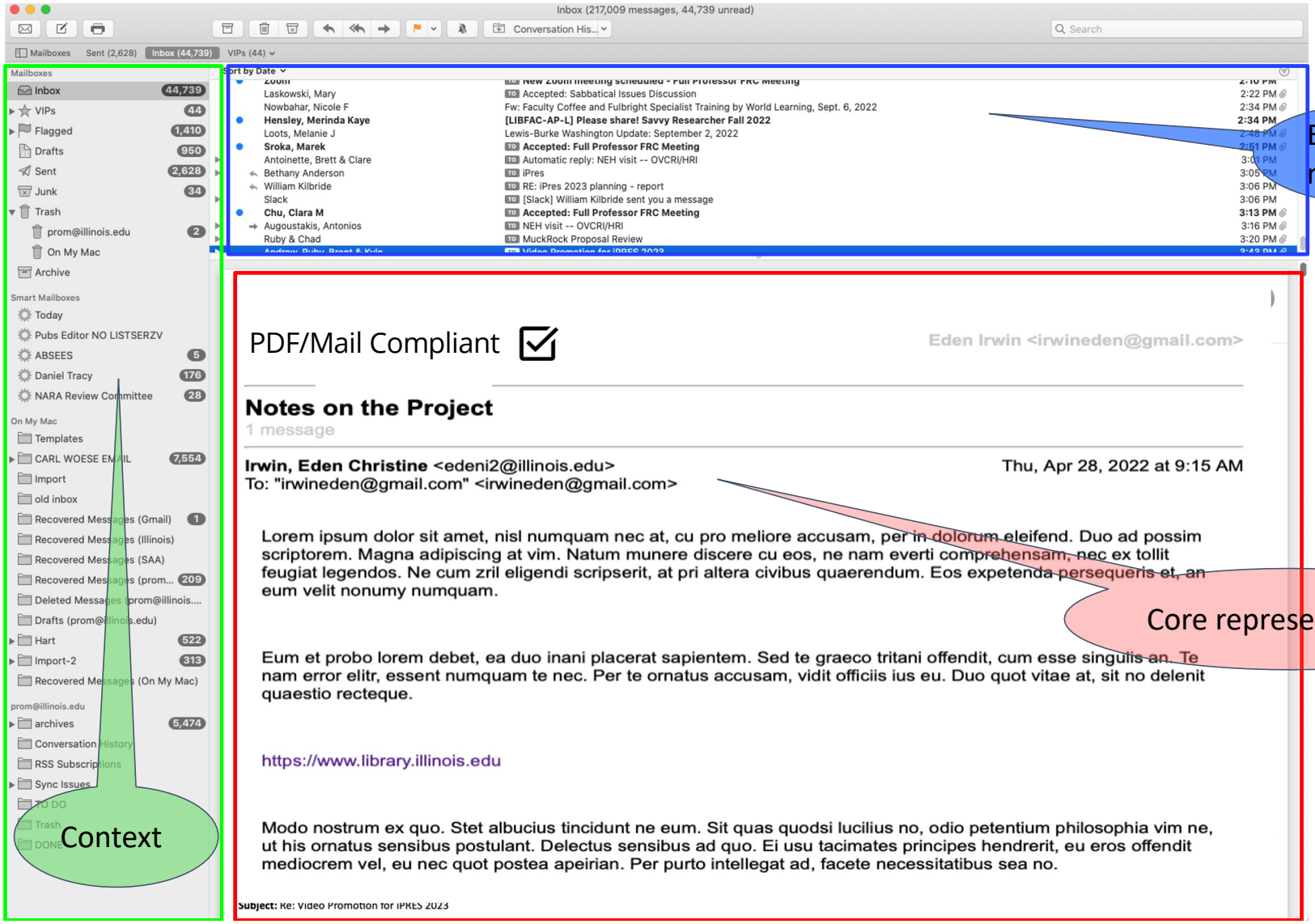
- Work-in-progress (v 0.2 of spec was released last week)
- Command line utility to create EA-PDF documents
  - GUI interface possible in the future
- Input is a folder of MBOX or EML files
  - PST possible in future
- Output is one or more EA-PDF documents
- Inspired by the DArcMail tool and the EAXS XML schema
- Open source
  - https://github.com/UIUCLibrary/ea-pdf
- High-level process flow on next slide

Process Flow for Conversion of MBOX to Archival PDF (EA-PDF)



Some Challenges

- Formatting HTML content, especially ‘ancient’ and badly formatted HTML
- Dealing with non-western languages (CJK)



ILLINOIS
University Library

Questions and Discussion