Email Archiving in PDF

Chris Prom and Eden Irwin
University of Illinois at Urbana-Champaign
Peter Wyatt
PDF Association

PDF Days Europe 2022 | Berlin

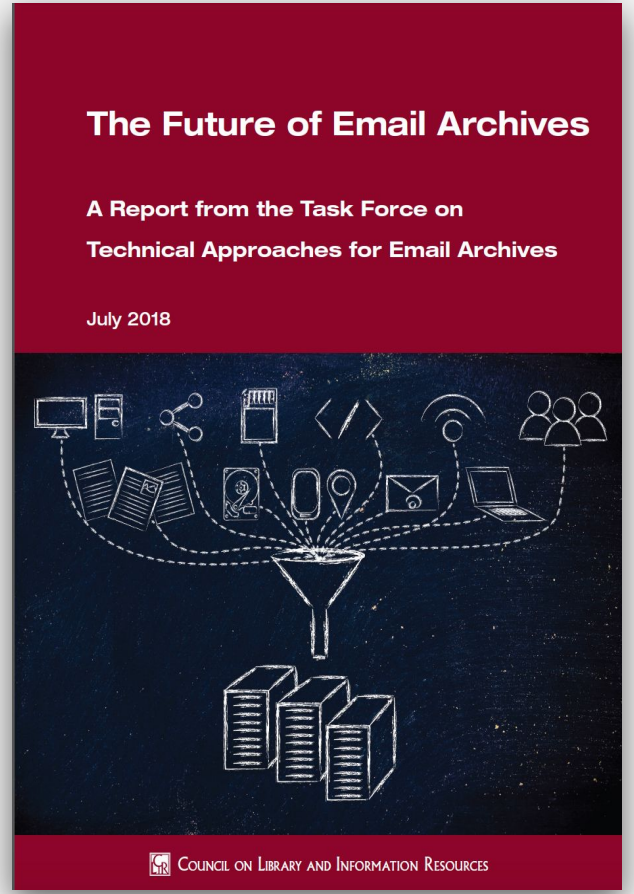
Why PDF/Mail Matters

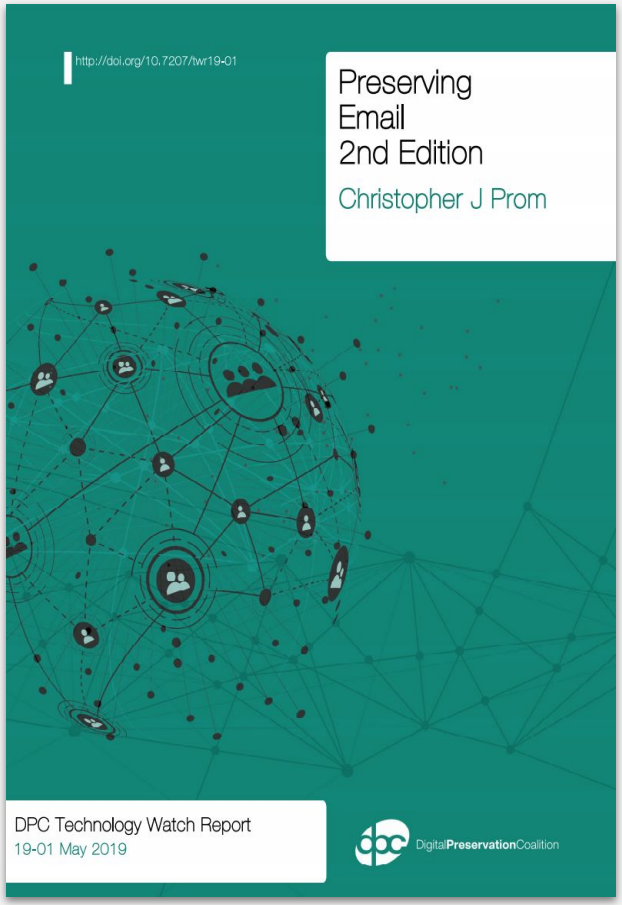
Chris:

- 10 min – need for PDF/Mail → may become PDF/M if and when ISO
- 1 min – Phase 1 Project Overview
- 4 min – Goals for Phase 2

Peter: 15 min – PDF/Mail as a Technology

20-25 slides


The Future of Email Archives
A Report from the Task Force on
Technical Approaches for Email Archives
July 2018
Council on Library and Information Resources


http://doi.org/10.7207/twr19-01
Preserving
Email
2nd Edition
Christopher J Prom
DPC Technology Watch Report
19-01 May 2019
DigitalPreservationCoalition

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

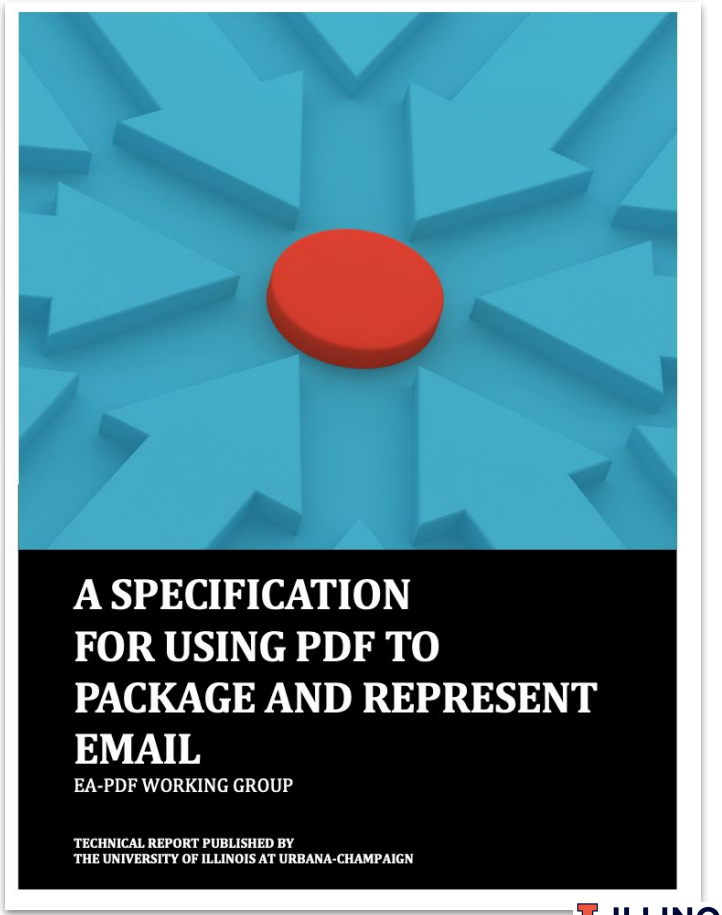
Matthew Hardy - Adobe Systems Inc.

Dietrich von Seggern, Callas software, GmbH

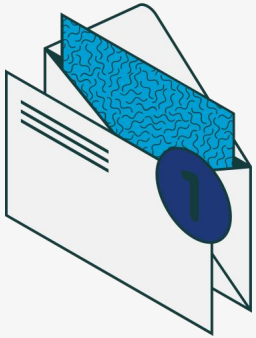
Joel Simpson - Artefactual Systems

Phase One Report

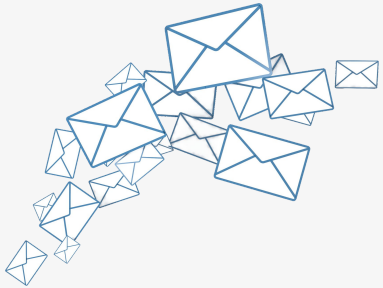
- Articulates rationale for EA-PDF (PDF/Mail)
- Define conceptual requirements for (PDF/Mail) container: a PDF file containing email data in defined structures and having several core archival attributes.
- Describe functional requirements for PDF/Mail-specific viewers.



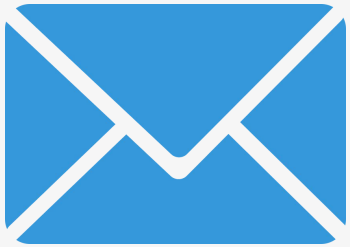
Why Preserve Email?



Email is one of the most ubiquitous forms of communication

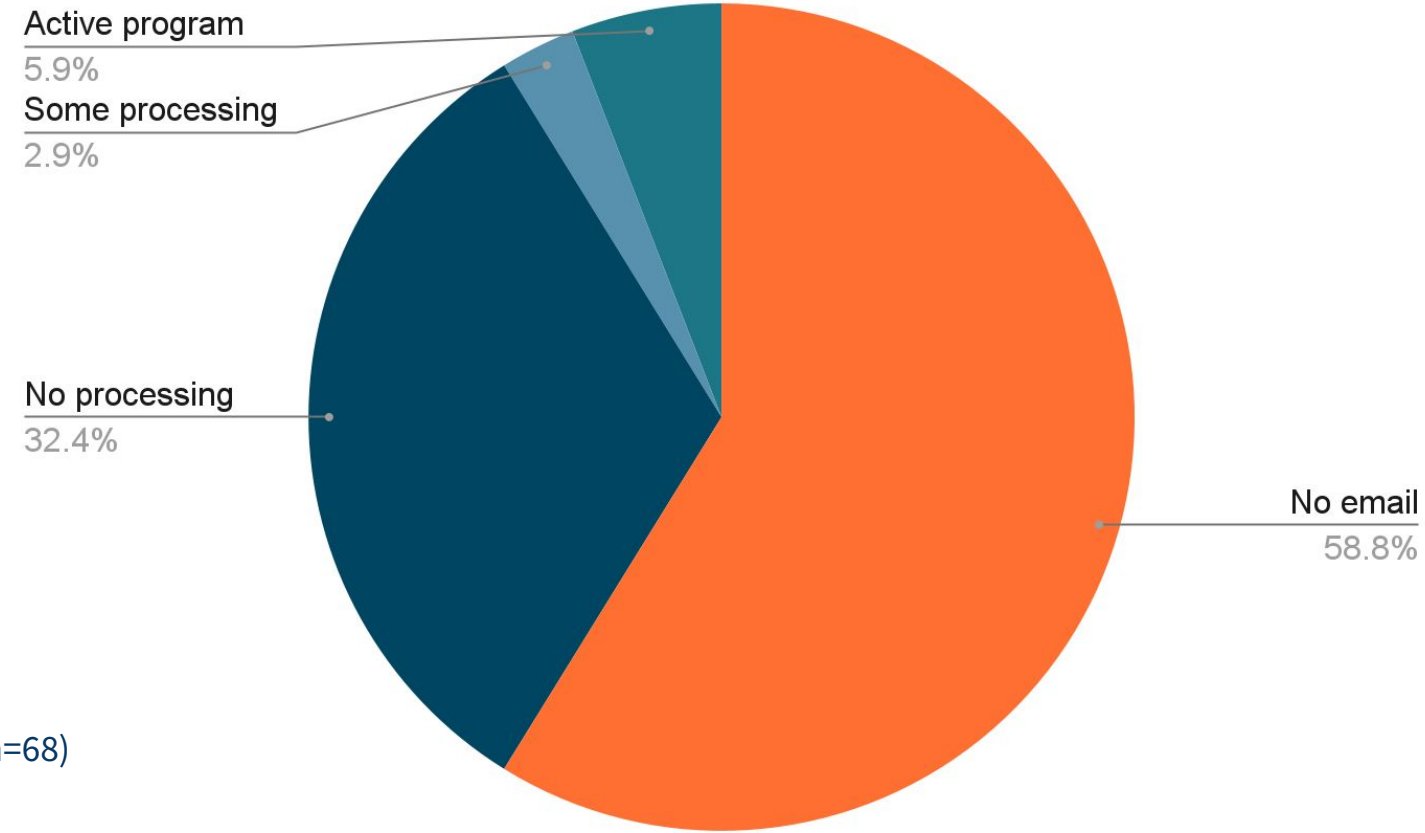


Email is historical evidence

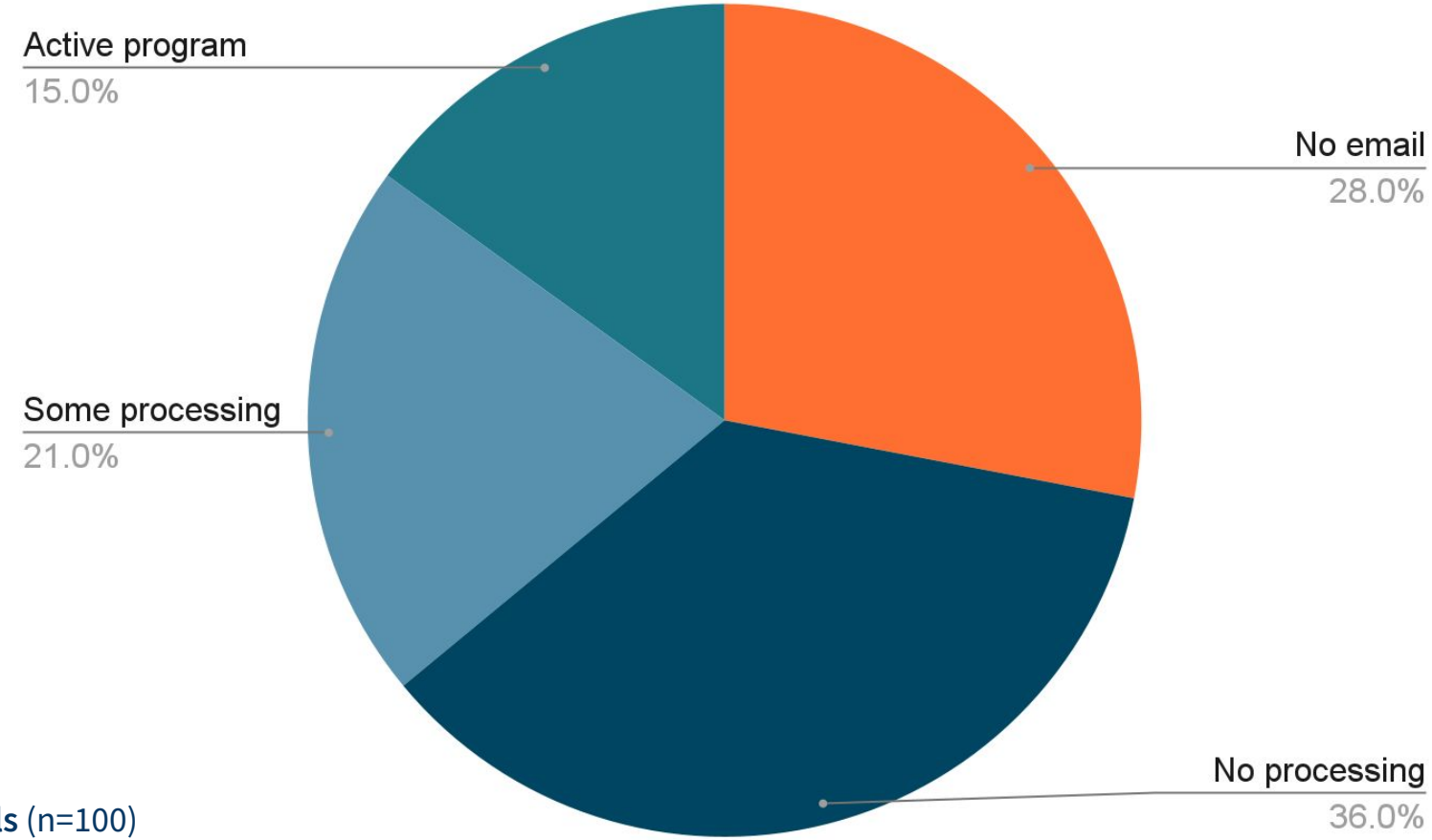


Email holds an incredible amount of information

Current State of Email Archiving (State of Illinois Sample)



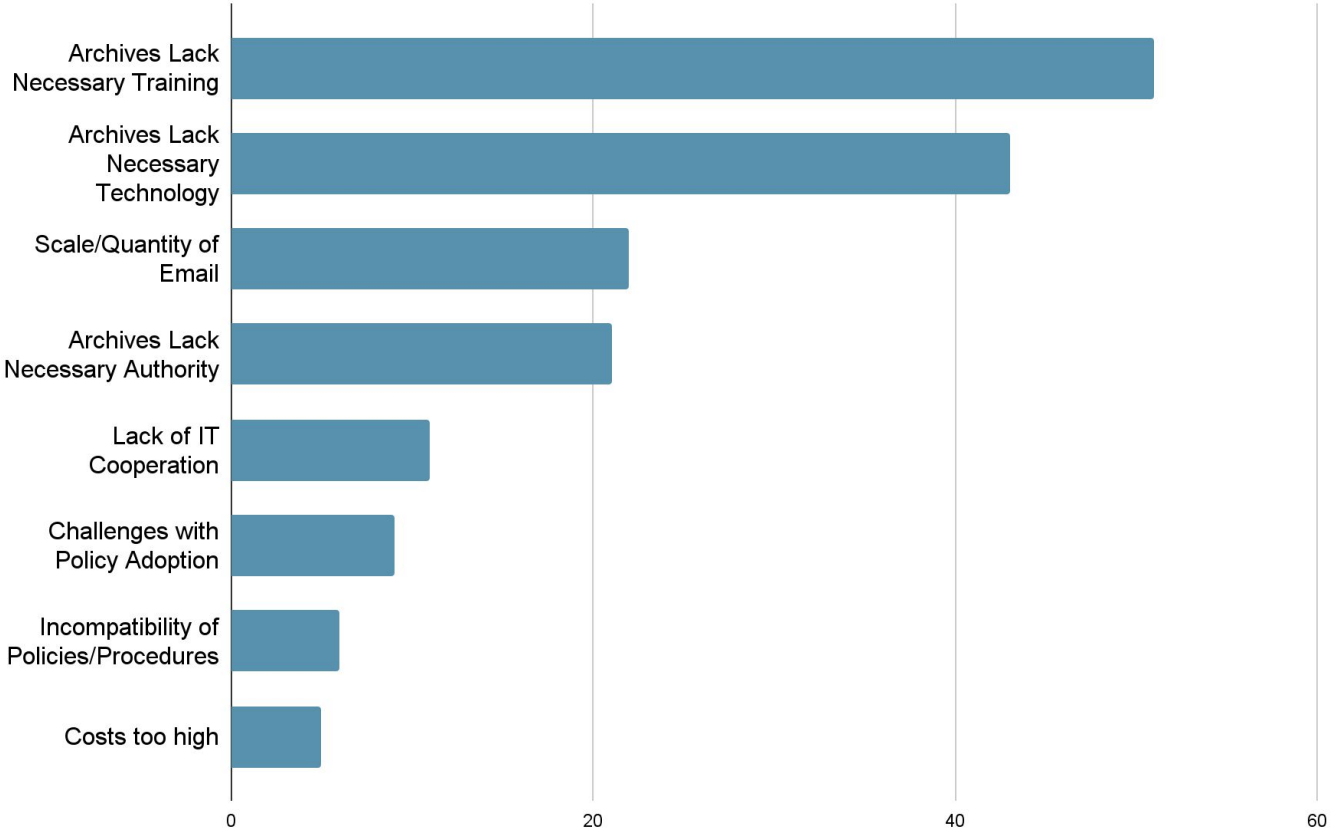
Survey (n=68)



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

Why PDF/Mail?

Complementary to preservation features of PDF/A

Standardizes metadata and evidential value

PDF is already used within archive community

Current PDF Functionality

Visible Message

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

Internal Information

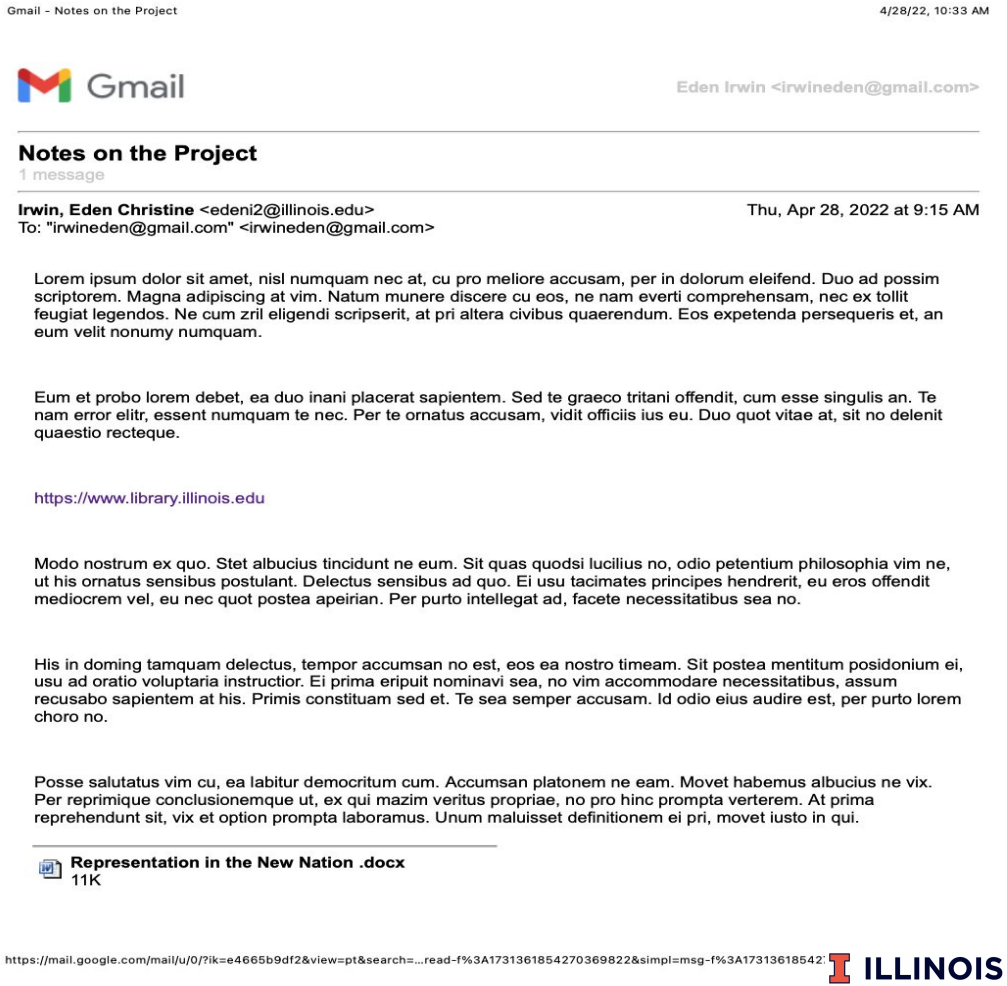
The only metadata that remains is about the created PDF, not the email

Attachments and Bulk Conversion

Neither can be expected

Links

Links remain active



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.

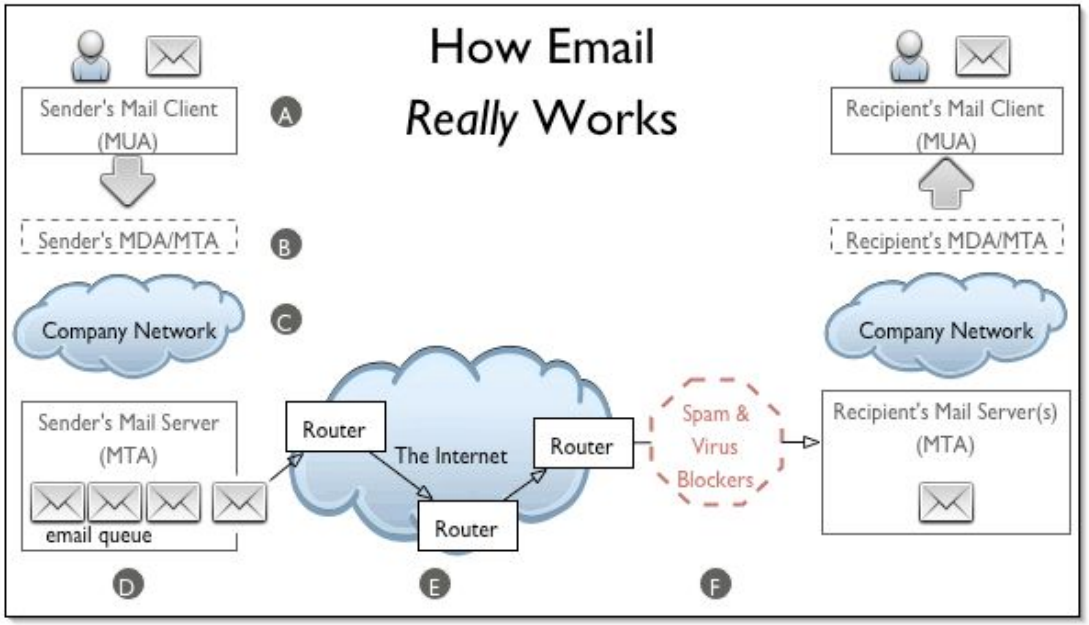
Posse salutatus vim cu, ea labitur democritum cum. Accumsan platonem ne eam. Movet habemus albucius ne vix. Per reprimique conclusionemque ut, ex qui mazim veritus propriae, no pro hinc prompta verterem. At prima reprehendunt sit, vix et option prompta laboramus. Unum maluisset definitionem ei pri, movet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

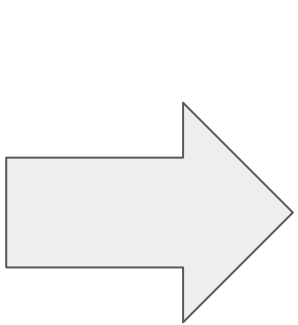
Representation in the New Nation .docx
11K

https://mail.google.com/mail/u/0/?ik=e4665b9df2&view=pt&search=...read-f%3A1731361854270369822&simpl=msg-f%3A17313618542...

ILLINOIS
University Library

Email Fixity for Archives





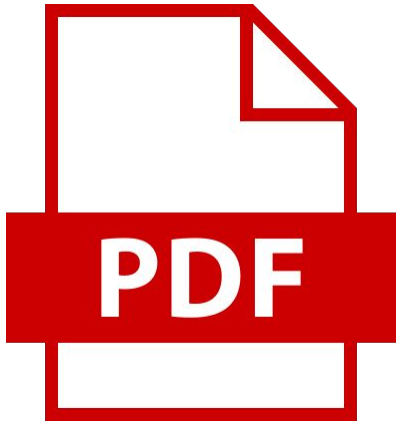


Image via: https://www.oasis-open.org/khelp/kmlm/user_help/html/how_email_works.html

Phase Two

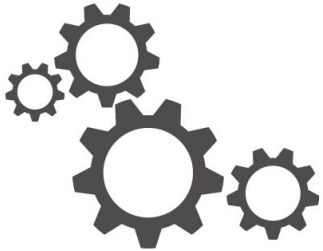
INSTITUTE of Museum and Library SERVICES

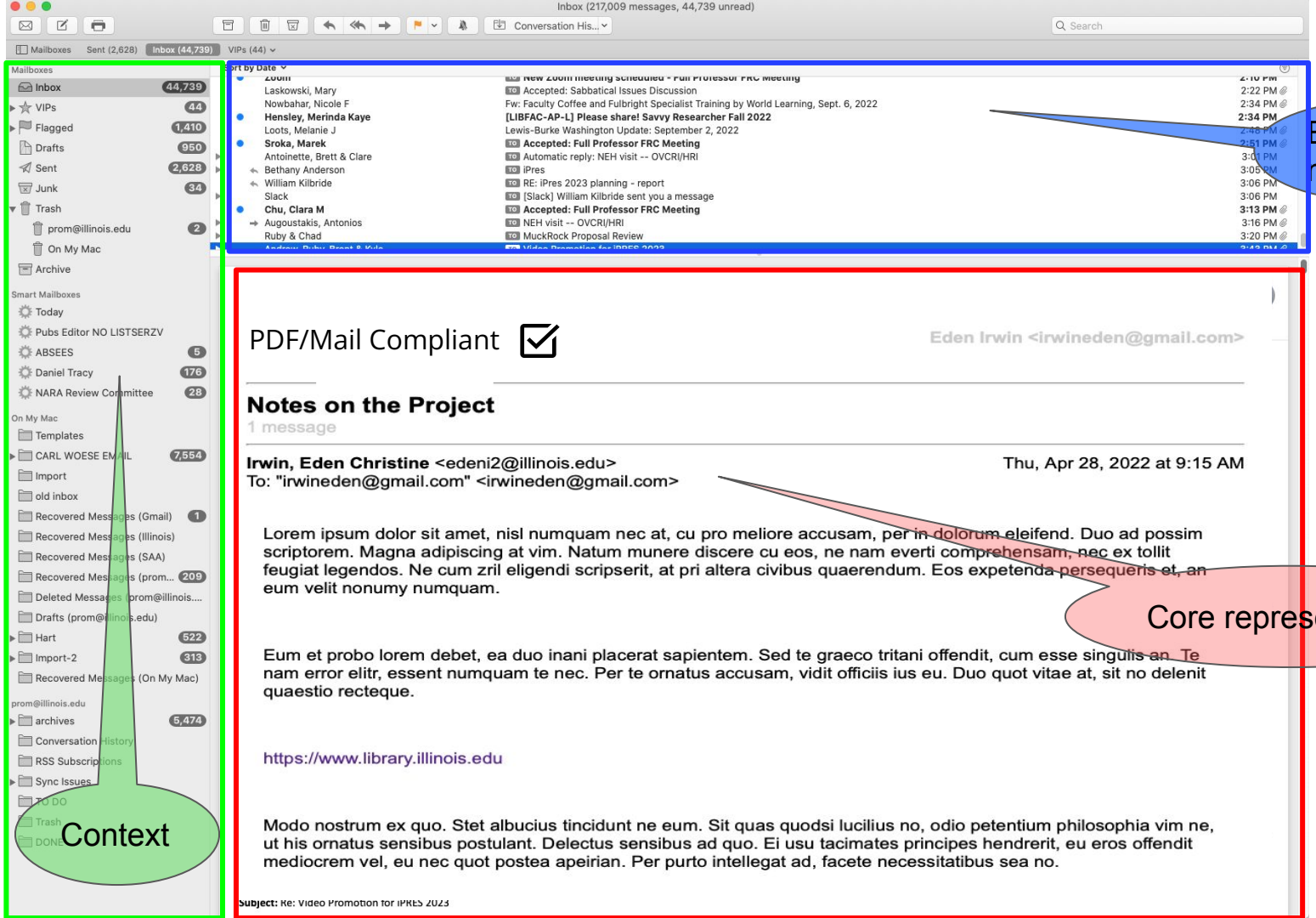
ILLINOIS University Library

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

NC DEPARTMENT OF
NATURAL AND CULTURAL RESOURCES

ILLINOIS
University Library

LIBRARY
LIBRARY OF CONGRESS

Williams

Association of Tribal
ATALM
Archives, Libraries, & Museums

IDAHO
TRANSPORTATION DEPARTMENT

UNIVERSITY AT ALBANY
State University of New York

McKESSON

MUCKROCK

VE RI TAS
HARVARD
UNIVERSITY

BROOKLYN NAVY YARD

Smithsonian

THE UNIVERSITY OF CALIFORNIA
LET THERE BE LIGHT
1868

NATIONAL ARCHIVES

ALMA MATER STUDIORUM
UNIVERSITAS DI BOLOGNA
A.D. 1088

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

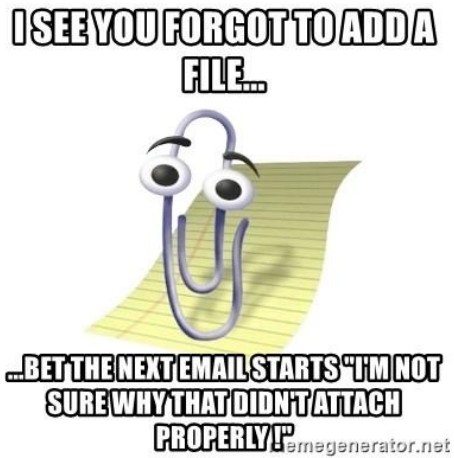
EA-PDF Technology

- **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)

Beyond “Core Representation” - key features of PDF/mail

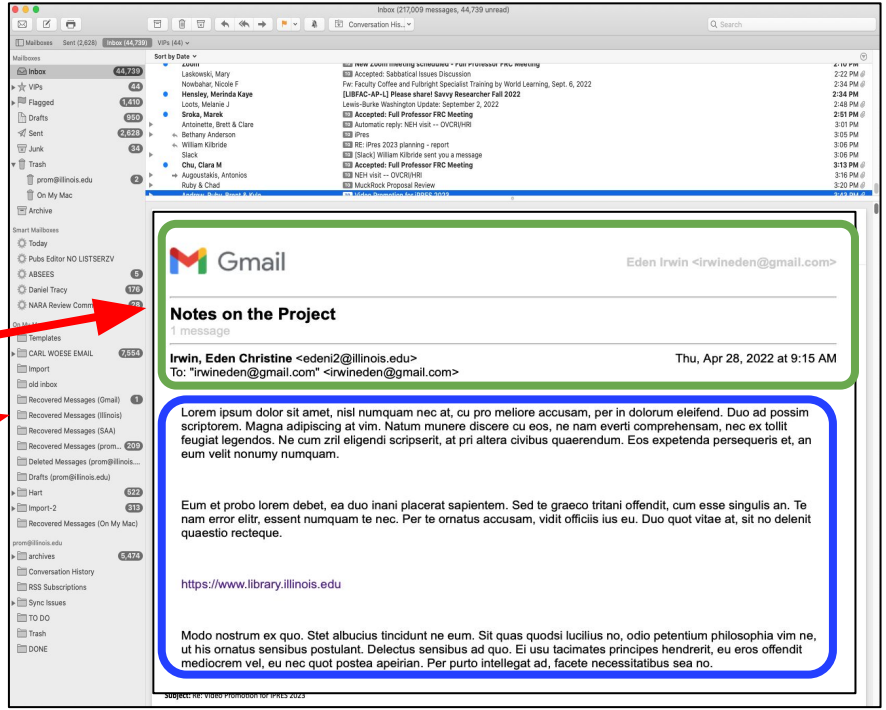
- XMP-based metadata
  - Namespace: `pdfmail`
- Embedded files (compressed)
  - The “raw” email file (EML, MBOX, PST, …)
  - Email attachments
  - HTML assets
- Associated Files and AFRelationship
  - At document and object level
- Navigational aids
  - Outlines, file attachment annotations

PREMIS
PRESERVATION METADATA
MAINTENANCE ACTIVITY



PDF/mail challenges

- Continuous → paginated
  - Knowing the source of a page
  - DPart / DPM
- Page appearance
- Hierarchical folder structure
  - PDF Portable collections
- Semantics of text-only email
- File size
  - Compression filters

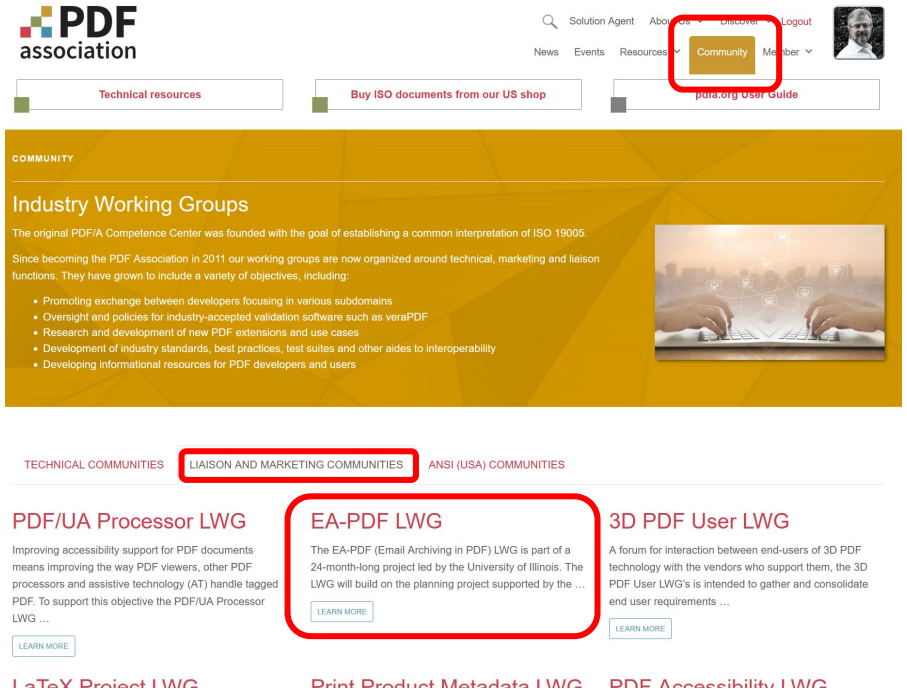


How to participate

PDF Association members can join the **EA-PDF Liaison Working Group** via the Member Area.

Non-members with expertise or professional interest in email technology can email info@pdfa.org and request to join.

Agreement to abide by the PDF Association's IPR policy and Code of Conduct is required for all working group members.



Questions?

Chris Prom and Eden Irwin
University of Illinois at Urbana-Champaign

Peter Wyatt
PDF Association