

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



Digital Preservation Coalition

Phase One

Support from Andrew W. Mellon Foundation

Chris Prom, University of Illinois (PI)

Kevin De Vorse - 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

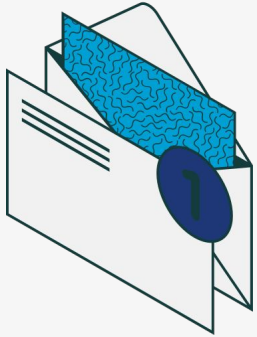
Joel Simpson - Artefactual Systems

Phase One Report

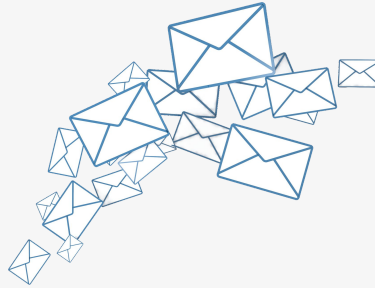
- Articulates rationale for EA-PDF (PDF/Mail)
- Define conceptual requirements for (PDF/Mail) container: a PDF file containing email data in defined structures and having several core archival attributes.
- Describe functional requirements for PDF/Mail-specific viewers.



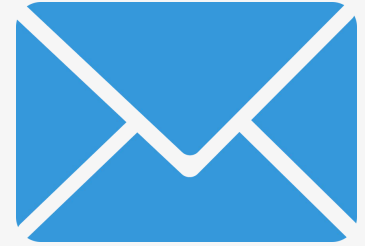
Why Preserve Email?



Email is one of the
most ubiquitous forms
of communication

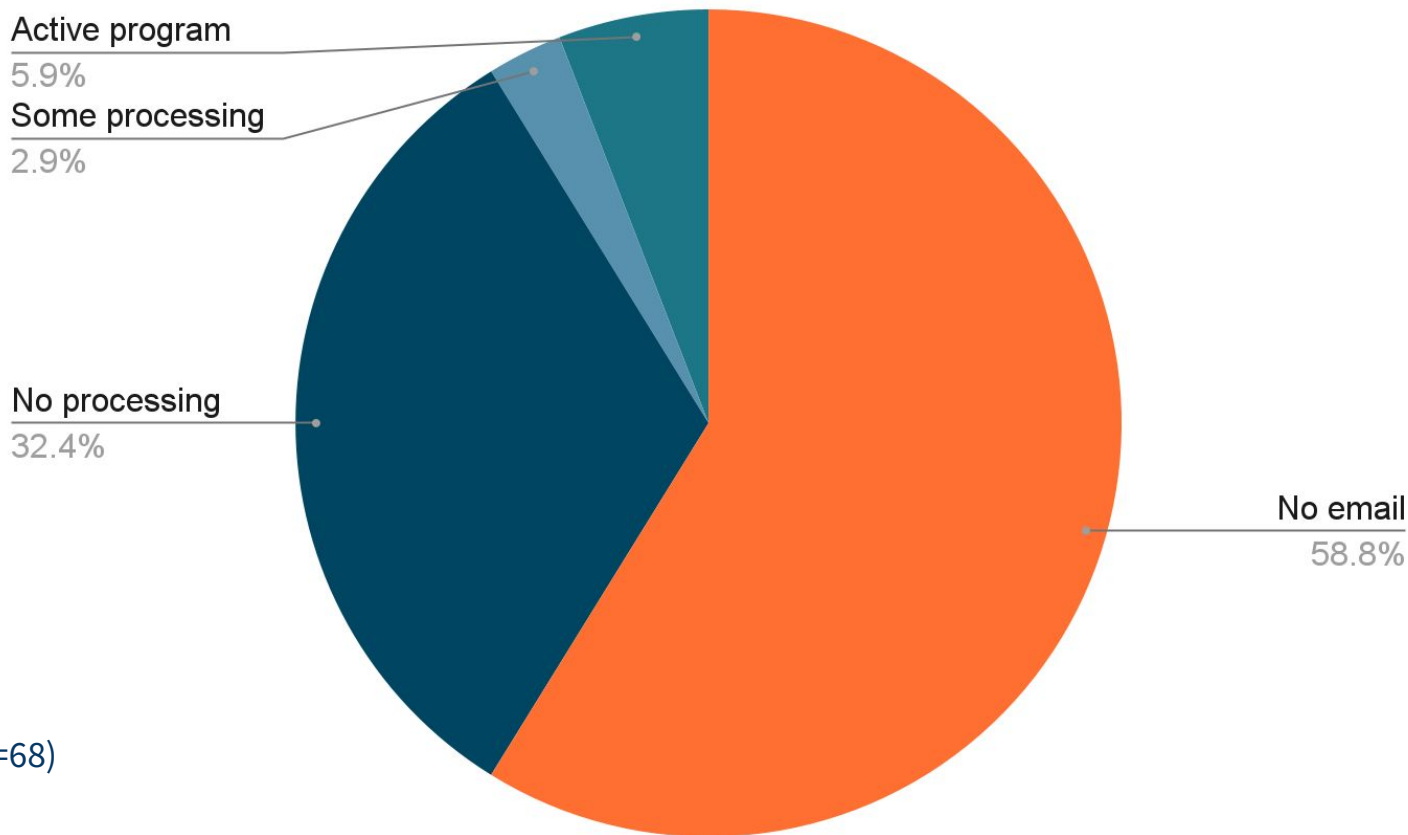


Email is historical
evidence



Email holds an
incredible amount of
information

Current State of Email Archiving (State of Illinois Sample)



Survey (n=68)

Active program

15.0%

No email

28.0%

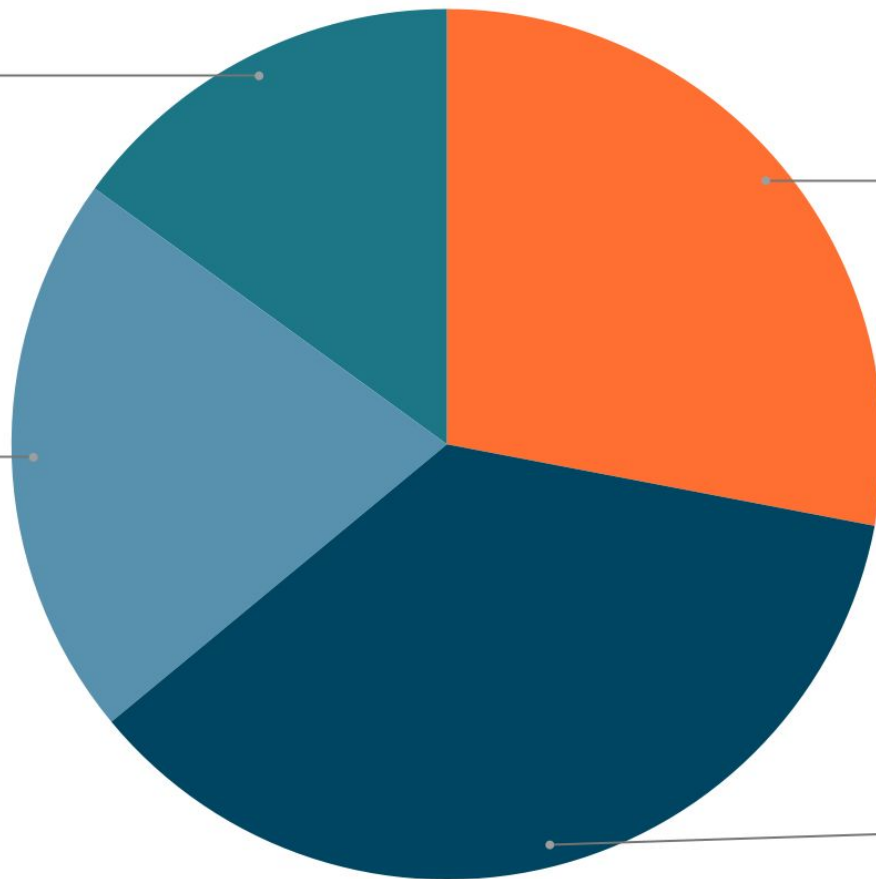
Some processing

21.0%

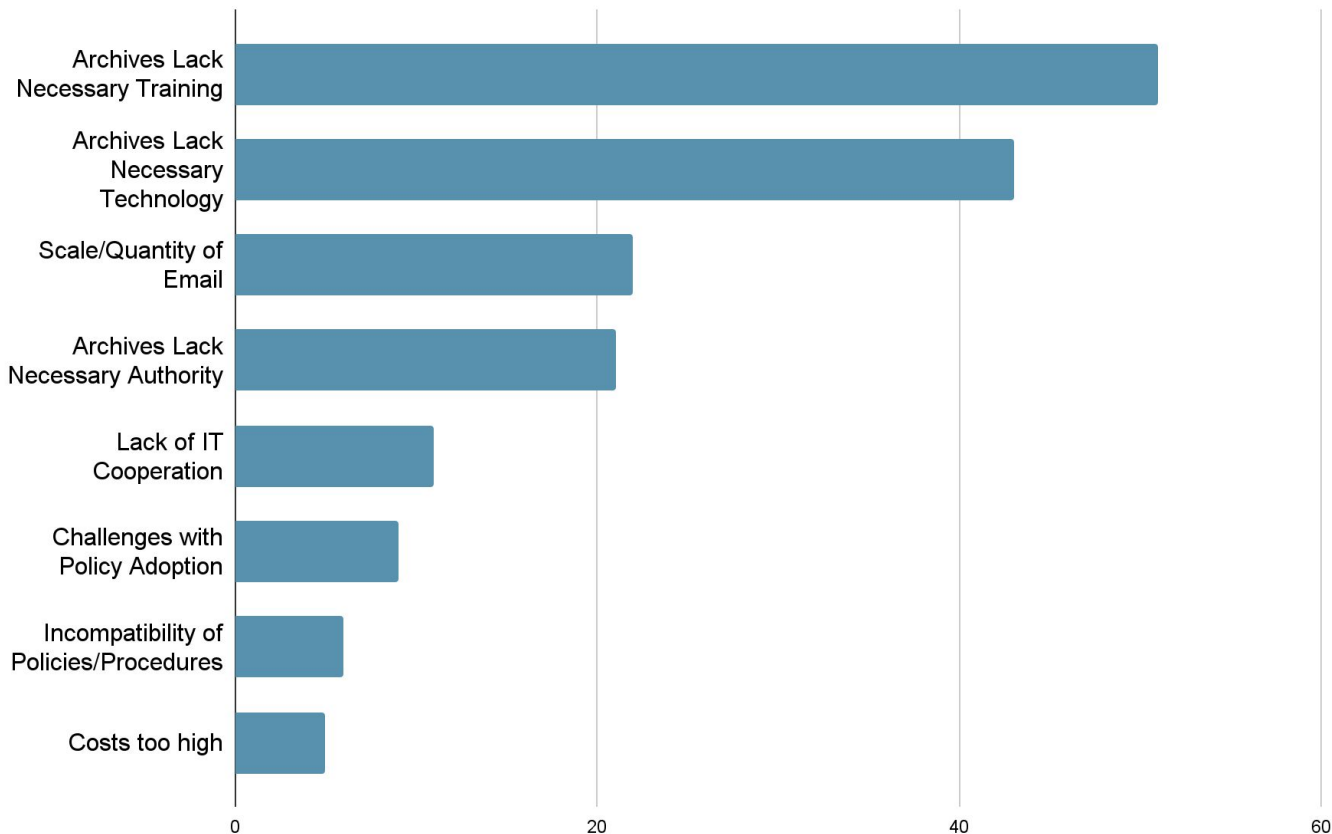
No processing

36.0%

Live Polls (n=100)



Why are archives not currently preserving email?



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



Eden Irwin <irwineden@gmail.com>

Notes on the Project

1 message

Irwin, Eden Christine <edeni2@illinois.edu>
To: "irwineden@gmail.com" <irwineden@gmail.com>

Thu, Apr 28, 2022 at 9:15 AM

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 elit, 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 albus 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 henderit, 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 mentium posidonium ei, usu ad oratio voluptaria instructor. 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 salutat vim cu, ea labitur democritum cum. Accumsan platonem ne eam. Movet habemus albus 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

Email Fixity for Archives

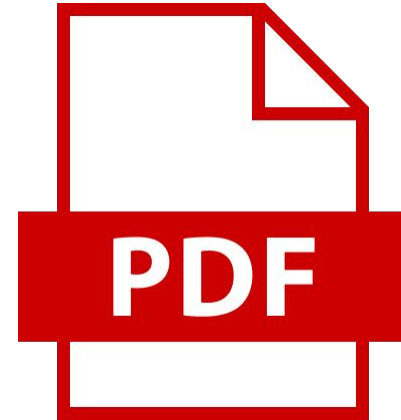
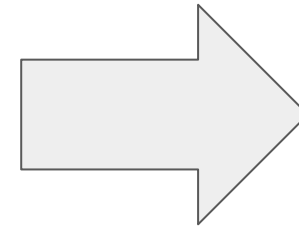
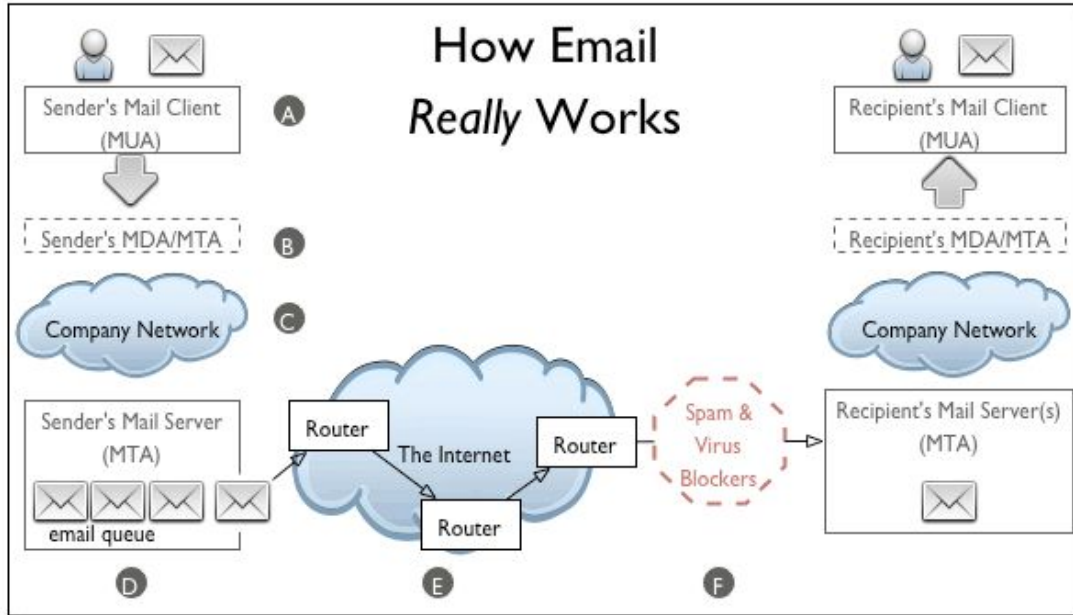


Image via: https://www.oasis-open.org/khelp/kmlm/user_help/html/how_email_works.html

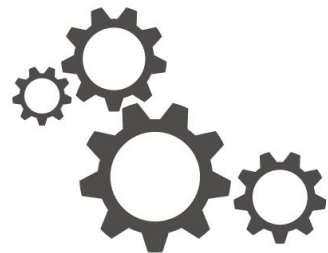
Phase Two

I ILLINOIS
University Library



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



Mailboxes Sent (2,628) Inbox (44,739) VIPs (44)

Mailboxes

- Inbox (44,739)
- VIPs (44)
- Flagged (1,410)
- Drafts (950)
- Sent (2,628)
- Junk (34)
- Trash
- On My Mac (2)
- Archive

Smart Mailboxes

- Today
- Pubs Editor NO LISTSERV (5)
- ABSEES
- Daniel Tracy (176)
- NARA Review Committee (28)

On My Mac

- Templates
- CARL WOESE EMAIL (7,554)
- Import
- old inbox
- Recovered Messages (Gmail) (1)
- Recovered Messages (Illinois)
- Recovered Messages (SAA)
- Recovered Messages (prom...) (209)
- Deleted Messages (prom@illinois...)
- Drafts (prom@illinois.edu)
- Hart (522)
- Import-2 (313)
- Recovered Messages (On My Mac)

prom@illinois.edu (5,474)

- archives
- Conversation History
- RSS Subscriptions
- Sync Issues
- TO DO
- Trash
- DONE

new zoom meeting scheduled - Full Professor FRC Meeting

Accepted: Sabbatical Issues Discussion

Fw: Faculty Coffee and Fulbright Specialist Training by World Learning, Sept. 6, 2022

[LIBFAC-AP-L] Please share! Savvy Researcher Fall 2022

Lewis-Burke Washington Update: September 2, 2022

Accepted: Full Professor FRC Meeting

Automatic reply: NEH visit -- OVCRI/HRI

iPres

RE: iPres 2023 planning - report

[Slack] William Kilbride sent you a message

Accepted: Full Professor FRC Meeting

NEH visit -- OVCRI/HRI

MuckRock Proposal Review

Video Promotion for IPRES 2023

EA-PDF
metadata

PDF/Mail Compliant ☒

Eden Irwin <irwineden@gmail.com>

Notes on the Project

1 message

Irwin, Eden Christine <edeni2@illinois.edu>
To: "irwineden@gmail.com" <irwineden@gmail.com>

Thu, Apr 28, 2022 at 9:15 AM

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 scriperit, at pri altera civibus quaerendum. Eos expetenda persequeris et, an eum velit nonumy numquam.

Core representation

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 albus 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 henderit, eu eros offendit mediocrem vel, eu nec quot postea apeirian. Per purto intellegat ad, facete necessitatibus sea no.

Subject: Re: Video Promotion for IPRES 2023

Context



NC DEPARTMENT OF
NATURAL AND CULTURAL RESOURCES

I ILLINOIS
University Library

LIBRARY
LIBRARY
OF CONGRESS

Williams



Association of Tribal



Archives, Libraries, & Museums



UNIVERSITY
AT ALBANY

State University of New York

McKESSON

MUCKROCK



HARVARD
UNIVERSITY



NATIONAL
ARCHIVES



BROOKLYN
NAVY YARD



Smithsonian



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)



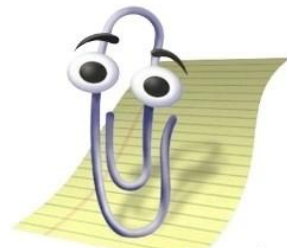
Beyond “Core Representation” - key features of PDF/mail

- 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

P R E M I S

PRESERVATION METADATA
MAINTENANCE ACTIVITY

I SEE YOU FORGOT TO ADD A
FILE...



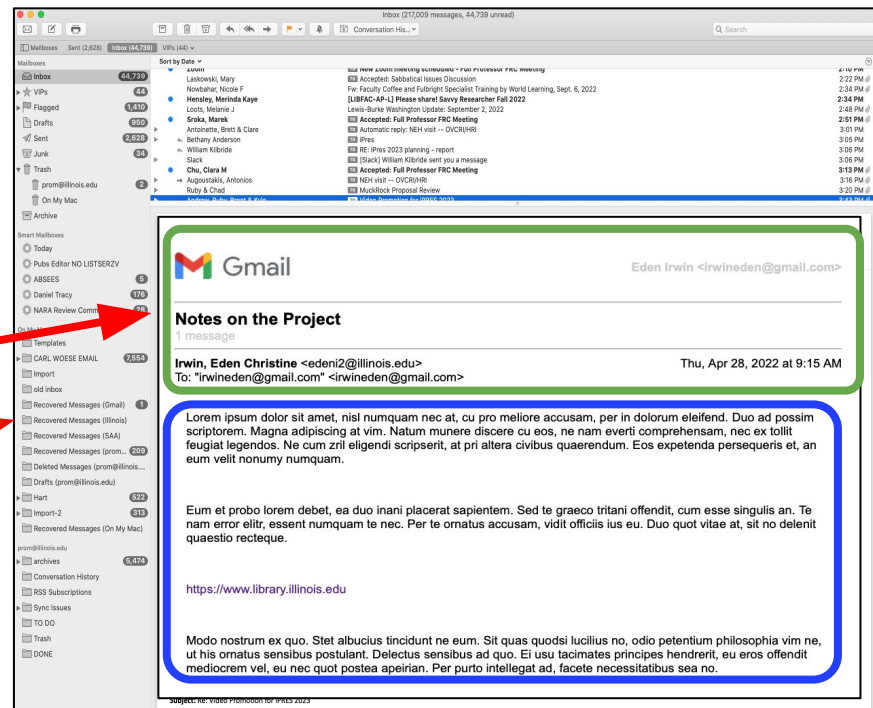
...BET THE NEXT EMAIL STARTS "I'M NOT
SURE WHY THAT DIDN'T ATTACH
PROPERLY!"

memegenerator.net



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.

The screenshot shows the PDF Association website. The top navigation bar includes links for Solution Agent, About, Discover, and Logout. Below this, there are links for News, Events, Resources, and a highlighted 'Community' link. A 'Buy ISO documents from our US shop' button is also visible. The main content area is titled 'COMMUNITY' and features a section for 'Industry Working Groups'. This section describes the PDF/A Competence Center's goal and lists objectives such as promoting exchange, oversight, research, and development of standards. A list of working groups is provided, with 'EA-PDF LWG' highlighted in a red box. The description for EA-PDF LWG states it is part of a 24-month-long project led by the University of Illinois. Other working groups listed include PDF/UA Processor LWG, 3D PDF User LWG, LaTeX Project LWG, Print Product Metadata LWG, and PDF Accessibility LWG.

PDF association

Solution Agent About Discover Logout

News Events Resources Community Member

Technical resources Buy ISO documents from our US shop pdfa.org User Guide

COMMUNITY

Industry Working Groups

The original PDF/A Competence Center was founded with the goal of establishing a common interpretation of ISO 19005.

Since becoming the PDF Association in 2011 our working groups are now organized around technical, marketing and liaison functions. They have grown to include a variety of objectives, including:

- Promoting exchange between developers focusing in various subdomains
- Oversight and policies for industry-accepted validation software such as veraPDF
- Research and development of new PDF extensions and use cases
- Development of industry standards, best practices, test suites and other aides to interoperability
- Developing informational resources for PDF developers and users

TECHNICAL COMMUNITIES LIAISON AND MARKETING COMMUNITIES ANSI (USA) COMMUNITIES

PDF/UA Processor LWG

Improving accessibility support for PDF documents means improving the way PDF viewers, other PDF processors and assistive technology (AT) handle tagged PDF. To support this objective the PDF/UA Processor LWG ...

LEARN MORE

EA-PDF LWG

The EA-PDF (Email Archiving in PDF) LWG is part of a 24-month-long project led by the University of Illinois. The LWG will build on the planning project supported by the ...

LEARN MORE

3D PDF User LWG

A forum for interaction between end-users of 3D PDF technology with the vendors who support them, the 3D PDF User LWG's is intended to gather and consolidate end user requirements ...

LEARN MORE

LaTeX Project LWG Print Product Metadata LWG PDF Accessibility LWG

Questions?

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

Peter Wyatt
PDF Association

