ILLINOIS

THE ANDREW W. MELLON FOUNDATION

OctoberPDFest ONLINE

A Specification for using PDF to Package and Represent Email

Christopher Prom, University of Illinois
@chrisprom
prom@illinois.edu

About Me

Budding Historian and Anglophile

Accidental Archivist

Digital Preservation Student

Digital Strategies Leader

About the Archives, RM, DP Community

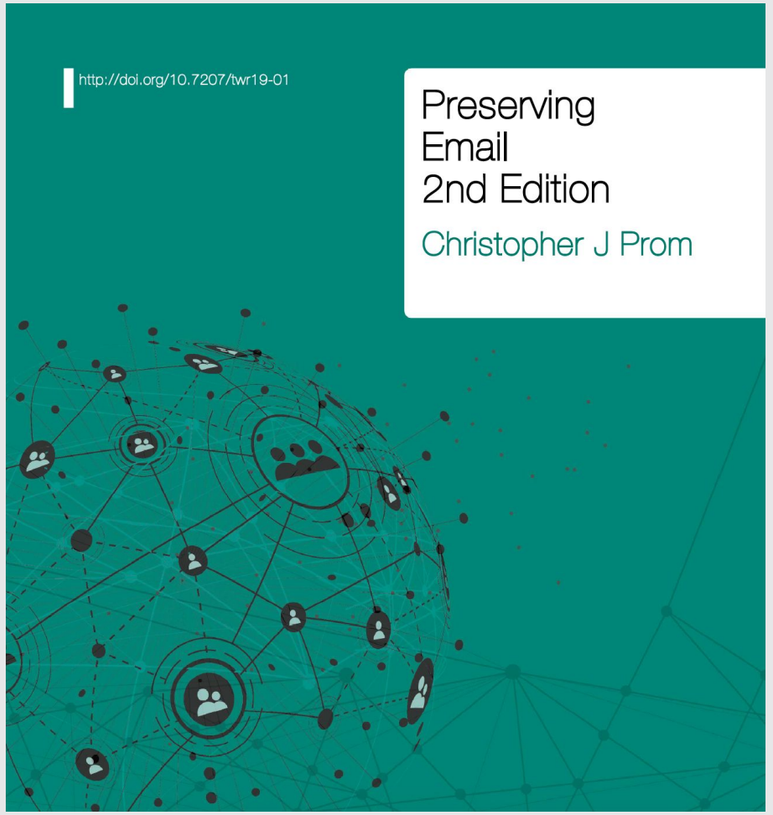
archivists.org arma.org dpcoline.org clir.org

Email Archiving Challenges

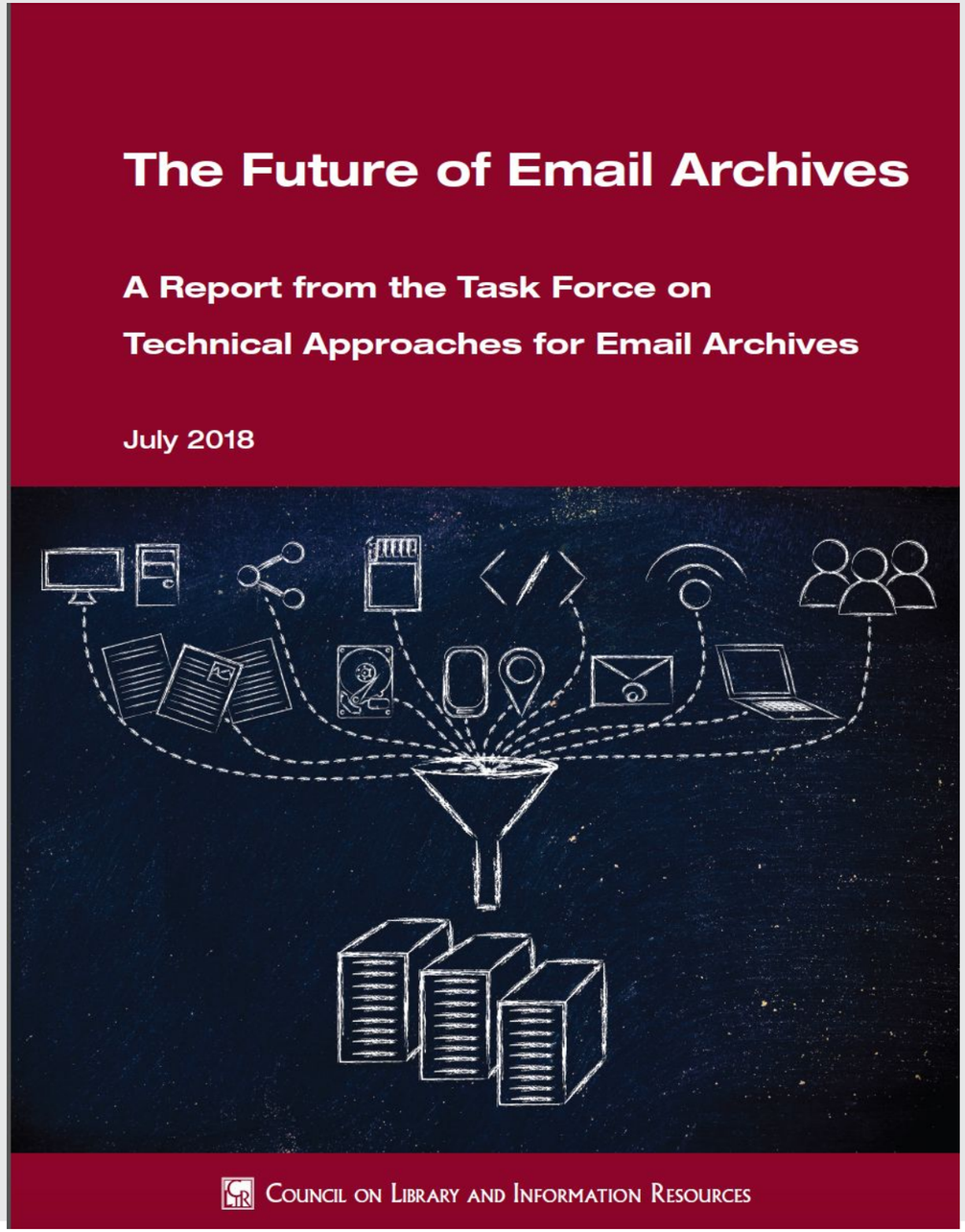
- How to capture?
  - “Appraisal,” privacy, sensitivity
  - Attachments, linked content
- How to preserve?
  - Preferred preservation/access formats?
  - File metadata support
  - Appropriate container formats
- How to scale?
  - Complex tool chaining
  - Reliance on community-specific tools

Deep Background/Extra Credit Reading

- Recommendations for information professionals who seek to preserve email for its cultural, legal or administrative value.
- Guidance to private individuals who may wish to preserve their email correspondence
- Describes the key policies, implementation strategies, procedures, tools, and services that can be drawn upon when developing an email preservation programme.
- Advocates for implementing appropriate technical standards, capture methods, and processing tools so that community can take practical steps to preserve email for its legal, administrative, or historical value.



http://doi.org/10.7207/twr19-01



THE ANDREW W. MELLON FOUNDATION

DigitalPreservationCoalition

www.clir.org/pubs/reports/pub175/

Email to PDF Working Group Members

Chris Prom, University of Illinois (PI)

Kevin De Vorsey - National Archives and Records Admin

Kate Murray - Library of Congress

Duff Johnson - PDF Association

Lynda Schmitz Fuhrig - Smithsonian Archives

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

Stephen Abrams - Harvard University Libraries

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

Matthew Hardy - Adobe Systems Inc.

Dietrich von Seggern, Callas software, GmbH

Joel Simpson - Artefactual Systems

“High Impact” Recommendations

- Sustain email archiving community
- Specification planning for beginning-of-lifecycle email tools
- Develop criteria for email authenticity
- Improve standards documentation
- **Improve options for PDF in email archiving workflows**
- Demonstrate value for email as research data source/data challenge

- Improved tools for sensitivity review
- **Develop email self-archiving tools**
- **Sustaining and integrating existing tools**
- **Develop standards for tool interoperability with a reference implementation**
- Machine learning and natural language processing
- Email archiving at scale

Specification Purpose

- Articulate the rationale for transforming email messages, folders, and accounts into archive-ready PDF packages
- Describe conceptual requirements for packaging one or many email messages into an “email archive using PDF” (EA-PDF) container: a PDF file containing email data in defined structures and having several core archival attributes.
- Describes functional requirements for EA-PDF-specific viewers.

https://go.illinois.edu/DraftEmailSpec

Why Convert Email to PDF???

Community Demand

xxxx xxxxx: I'm very excited about this work. It will be helpful in organizations for supporting retention schedules. I'm using print to PDF now for the most part and am a bit uneasy about it.

xxxxxxx xxxxxxxxxx: Seems like PDF would be a great low-tech bar of entry for users to access email archives.

xxxx xxxxxx It will be a great benefit to lone arrangers with limited IT support provided the tools are GUI supported.

xxxxx xxxxxxxxxx-xxxxxx: Agree!

xxxxxxx xxxxxxxxxx: It would take so much one-on-one time for reading room staff to guide low-tech researchers through software. PDF would be more of a neutral meeting ground for that.

Tool Roster

OctoberPDFest ONLINE

FTK Imager

Aid4Mail™

archivematica®

BitCurator

RATOM

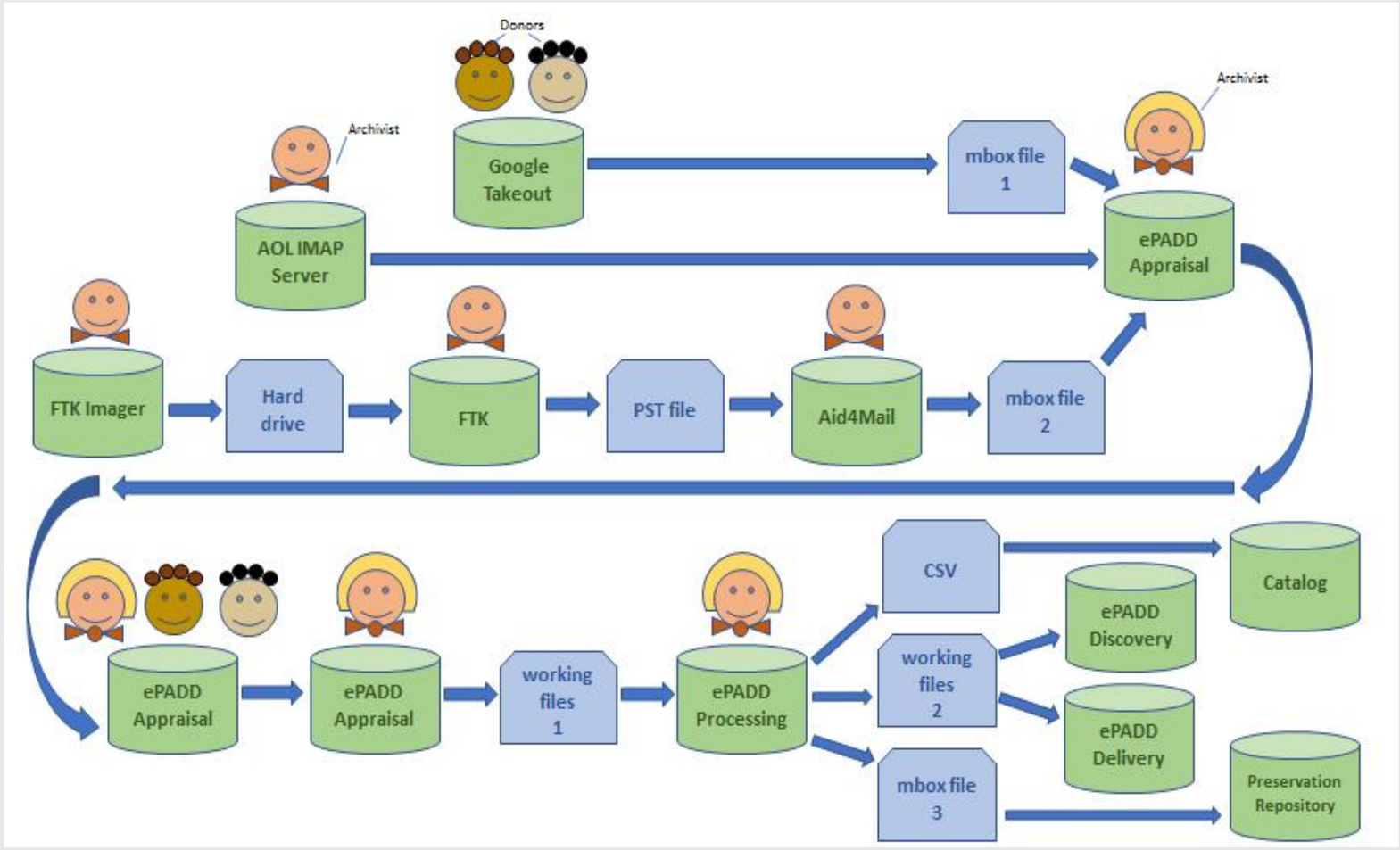
Review, Appraisal, and Triage of Mail

Smithsonian Institution Archives

ePADD

Transforming Online Mail with Embedded Semantics (TOMES)

Workflow scenario:



Do we need another standard for email?

The core email standards (IETF family of standards starting with RFC 5322) were designed for *sending* emails; not for storage, capture or later representation.

The most common email formats (.pst, .mbox etc.) in the vast majority of cases are not truly ‘original formats’ but exports taken at a point in time.

We believe those formats *can* be improved on, in two main ways:

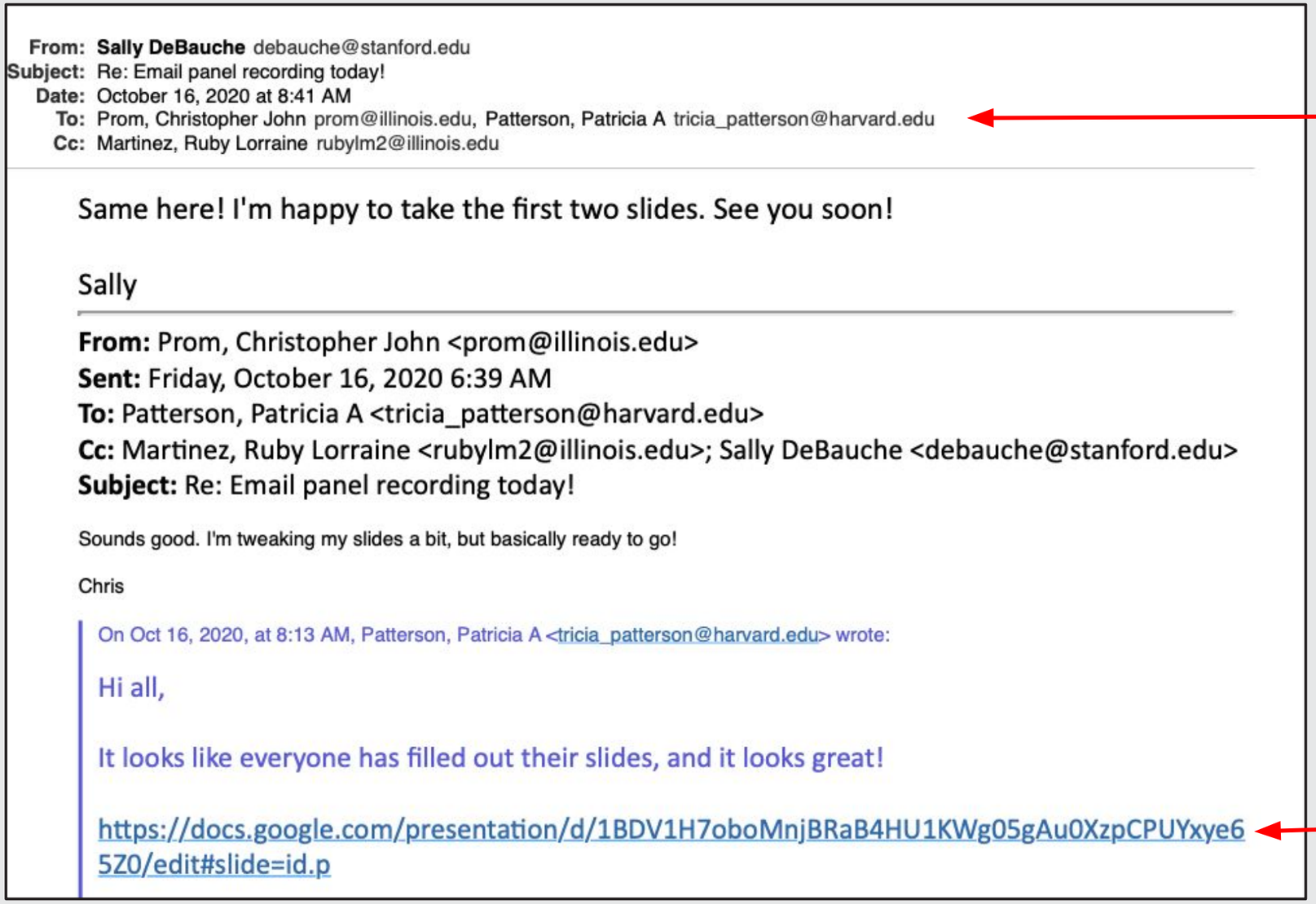
- standardizing the data-model for email capture
- capturing more information & context at the point of creation
- improving the quality and usability of exported email for the purposes of both preservation and access.

PDF has . . .

- *Existing market*
  - While current pdf tools are inadequate, the market already exists
- *Strong ecosystem*
  - Hundreds of PDF technology vendors; PDF basic viewing is almost universally supported
  - 160+ members of the PDF Association, which already supports software and standards development
- *Technical fit to requirements*
  - PDF 2.0 provides great support for technically complex containers, metadata, and associated content

Overview of Functional Specs

- **Open Standards:** Superset leveraging PDF 2.0 (ISO 32000-2); packages ‘should’ conform to ISO PDF/A and PDF/UA-2; structured metadata (XMP)
- **Capturing Email:** All messages, headers, body parts, and captured attachments by default; tools should support options to include exclude based on policy; options for linked content.
- **Describing Email:** Exclusions should be documented; EA-PDF package as whole should be described; provenance tracked; packages signed
- **Representing Email:** legacy reader support; core representation: collections defined; searchability
- **EA-PDF Viewer Functionality:** full support for core representation display properties and search; extraction of core representation as EML



From: **Sally DeBauche** debauche@stanford.edu
Subject: Re: Email panel recording today!
Date: October 16, 2020 at 8:41 AM
To: Prom, Christopher John prom@illinois.edu, Patterson, Patricia A tricia_patterson@harvard.edu
Cc: Martinez, Ruby Lorraine rubylm2@illinois.edu

Same here! I'm happy to take the first two slides. See you soon!

Sally

From: Prom, Christopher John <prom@illinois.edu>
Sent: Friday, October 16, 2020 6:39 AM
To: Patterson, Patricia A <tricia_patterson@harvard.edu>
Cc: Martinez, Ruby Lorraine <rubylm2@illinois.edu>; Sally DeBauche <debauche@stanford.edu>
Subject: Re: Email panel recording today!

Sounds good. I'm tweaking my slides a bit, but basically ready to go!

Chris

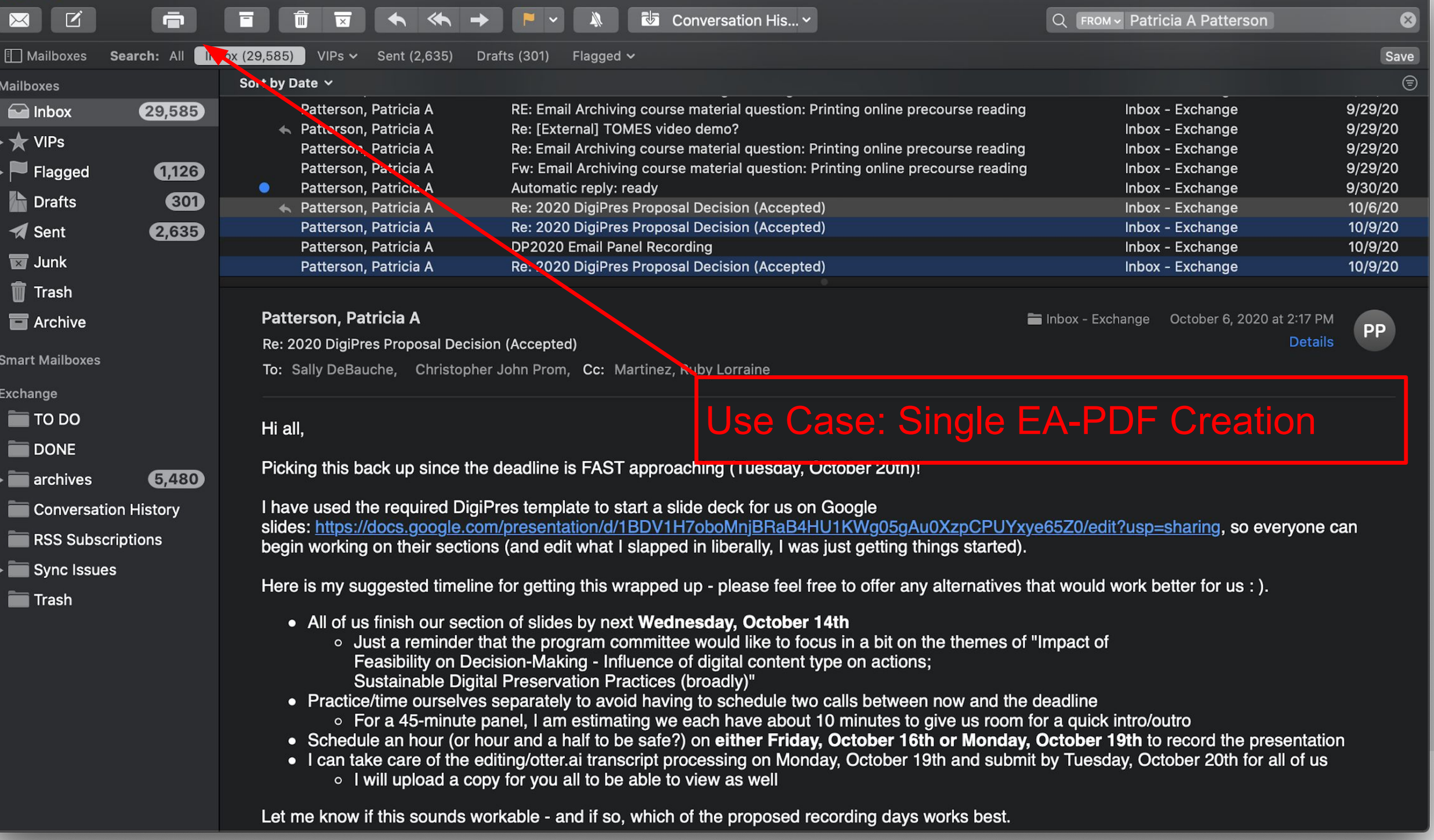
> On Oct 16, 2020, at 8:13 AM, Patterson, Patricia A <tricia_patterson@harvard.edu> wrote:
>
> Hi all,
>
> It looks like everyone has filled out their slides, and it looks great!
>
> https://docs.google.com/presentation/d/1BDV1H7oboMnjBRaB4HU1KWg05gAu0XzpCPUYxye65Z0/edit#slide=id.p

Limited Headers

Package Provenance?

Attachments?

Broken Link?


Conversation His...
FROM Patricia A Patterson
Mailboxes
Search: All
Inbox (29,585)
VIPs
Sent (2,635)
Drafts (301)
Flagged
Save
Sort by Date
Mailboxes
Inbox 29,585
VIPs
Flagged 1,126
Drafts 301
Sent 2,635
Junk
Trash
Archive
Smart Mailboxes
Exchange
TO DO
DONE
archives 5,480
Conversation History
RSS Subscriptions
Sync Issues
Trash
Patterson, Patricia A RE: Email Archiving course material question: Printing online precourse reading Inbox - Exchange 9/29/20
Patterson, Patricia A Re: [External] TOMES video demo? Inbox - Exchange 9/29/20
Patterson, Patricia A Re: Email Archiving course material question: Printing online precourse reading Inbox - Exchange 9/29/20
Patterson, Patricia A Fw: Email Archiving course material question: Printing online precourse reading Inbox - Exchange 9/29/20
Patterson, Patricia A Automatic reply: ready Inbox - Exchange 9/30/20
Patterson, Patricia A Re: 2020 DigiPres Proposal Decision (Accepted) Inbox - Exchange 10/6/20
Patterson, Patricia A Re: 2020 DigiPres Proposal Decision (Accepted) Inbox - Exchange 10/9/20
Patterson, Patricia A DP2020 Email Panel Recording Inbox - Exchange 10/9/20
Patterson, Patricia A Re: 2020 DigiPres Proposal Decision (Accepted) Inbox - Exchange 10/9/20
Patterson, Patricia A
Inbox - Exchange
October 6, 2020 at 2:17 PM
Details
PP
Re: 2020 DigiPres Proposal Decision (Accepted)
To: Sally DeBauche, Christopher John Prom, Cc: Martinez, Ruby Lorraine
Use Case: Single EA-PDF Creation
Hi all,
Picking this back up since the deadline is FAST approaching (Tuesday, October 20th)!
I have used the required DigiPres template to start a slide deck for us on Google slides: https://docs.google.com/presentation/d/1BDV1H7oboMnjBRaB4HU1KWg05gAu0XzpCPUYxye65Z0/edit?usp=sharing, so everyone can begin working on their sections (and edit what I slapped in liberally, I was just getting things started).
Here is my suggested timeline for getting this wrapped up - please feel free to offer any alternatives that would work better for us :).
All of us finish our section of slides by next Wednesday, October 14th
Just a reminder that the program committee would like to focus in a bit on the themes of "Impact of Feasibility on Decision-Making - Influence of digital content type on actions; Sustainable Digital Preservation Practices (broadly)"
Practice/time ourselves separately to avoid having to schedule two calls between now and the deadline
For a 45-minute panel, I am estimating we each have about 10 minutes to give us room for a quick intro/outro
Schedule an hour (or hour and a half to be safe?) on either Friday, October 16th or Monday, October 19th to record the presentation
I can take care of the editing/otter.ai transcript processing on Monday, October 19th and submit by Tuesday, October 20th for all of us
I will upload a copy for you all to be able to view as well
Let me know if this sounds workable - and if so, which of the proposed recording days works best.

What's Next? Proof of concept

- Draft the format specification based on the functional specification
- From an MBOX file, create EA-PDF for:
  - a single email message
  - a folder of messages
  - a message thread
  - an entire account/inbox
- An independent implementer will use the format specification to implement round trip conversion, exporting EA-PDF to EML/MBOX.
- Partnerships with PDF Association and members
- ISO standard for "EA-PDF" not contemplated at this time

Questions/Discussion