

Dietrich von Seggern

CEO
callas software

Vice Chair
PDF Association



Archiving email - as PDF?

The EA-PDF project



callas pdfaPilot PDF/A technology



- Since the standard exists (2005)
- In Adobe Acrobat, Foxit Phantom and other applications
- Includes email and office conversion
- Library (SDK), command line, server (hotfolder), desktop



The email to PDF/A reality show

Oct. 20 at 16:00 CEST / 10:00 ET / 07:00 PT

What of EA-PDF is already available (in a product)?



How PDF/A-2 and -4 are better than PDF/A-1

Oct. 21 at 1630 CEST / 1030 ET / 0730 PT

Why should you consider to update your archive format?



Intro

Email (archival) in a nutshell



- Essential part of today's business communication
- Business records documenting discussions, decisions, and actions
- Reasons for email archiving
 - Business needs of the organizations (industry, insurances, customer communication)
 - Legal requirements (archival of business records)
- But archiving email is not common - why is that so and how can we overcome that?

EA-PDF Specification (2021)



- EA-PDF: Email archiving in PDF
- Led by Archivists and academics, supported by PDF Association
- “Set the stage for software developers to create email capture and representation systems leveraging PDF”
- Phase 1 result: 34 pages, free for download (Funded by Andrew W. Mellon Foundation)



EA-PDF status - phase 2



- Funding for phase 2 granted
- Academic/industry partnership in an **EA-PDF Liaison Working Group** hosted by PDF Association
 - File format (ultimately an ISO subset standard?)
 - Implementer guidance for creation and viewing software
- Proof-of-concept, open source implementation



Poll

— Planning for phase 2 guidance

What describes your organizations email archiving policy best?

- ☐ We do not have a policy
- ☐ Individual responsibility of employees
- ☐ Central services archiving original email formats
- ☐ Central services archiving PDF
- ☐ Central services, I do not know the format
- Implementer guidance



The email archival file format?

Email archival - file format questions



- PDF/A has resolved the archival format dilemma for all paper like formats (no need to archive the viewer and proper environment for digital formats)
- For email often original formats are used for archiving
- What is the email format?

The “EML” format



- File representation in conformity with RFC 5322 (Internet Message Format, IMF)
- “There is no known specification that defines EML as a file format to store email messages on a file system.” (Library of Congress)
- And IMF is limited to text emails

But email is much more than text ...



- ...or even HTML
- Additional protocols (for interaction between email and other systems):
 - Email distribution lists
 - Encryption and digital signatures
 - Calendar invitations
 - ...
 - Winmail.dat...

Or “winmail.dat” attachments

An email in Apple mail has no body but a single attachment: winmail.dat



Apple mail vs. MS Outlook

The very same email - makes more sense when received with Outlook

The image shows a side-by-side comparison of the same email in two different email clients. On the left is the Apple Mail interface, and on the right is the Microsoft Outlook interface.

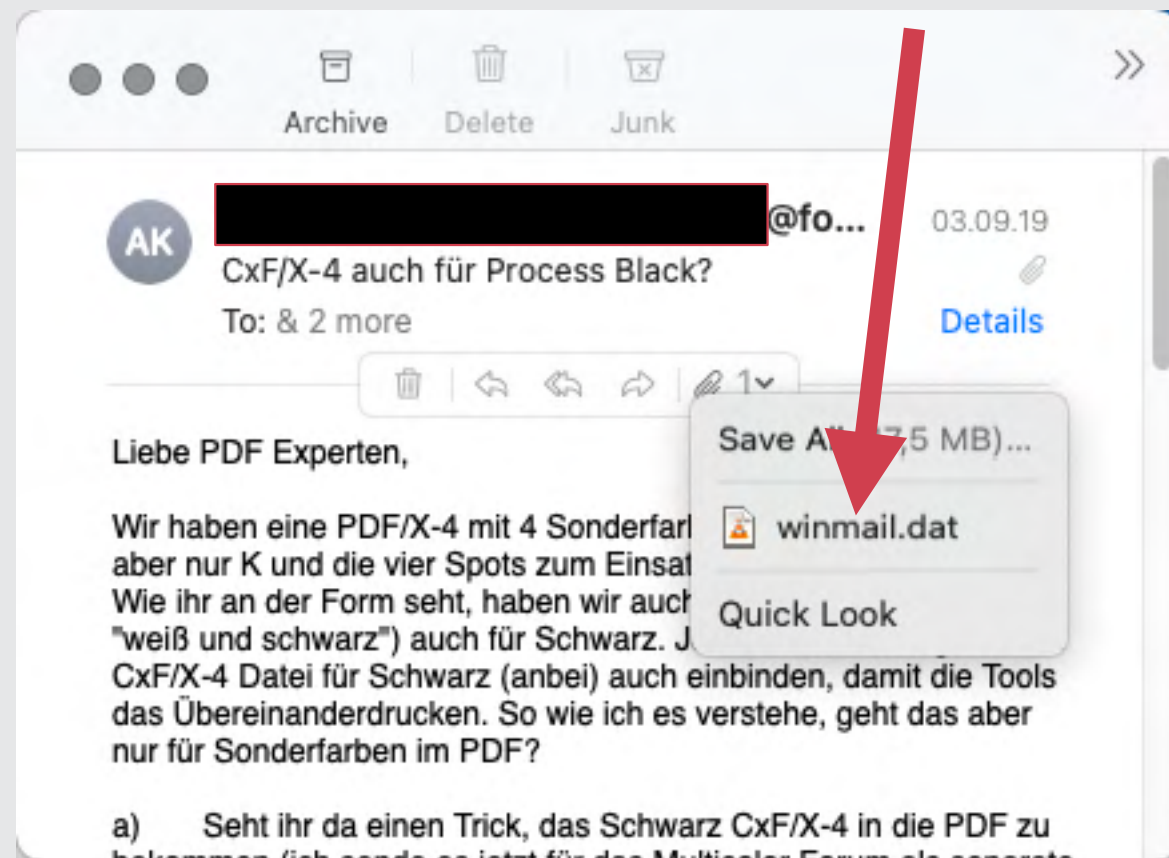
Apple Mail (Left): The email header shows the sender as 'immobilien+projektdienstleistungen' and the subject as 'Anforderung der Einzelerzeugung von pdf-Files'. The body of the email is partially visible, showing a greeting 'guten Tag' and a paragraph about electronic billing. A file named 'winmail.dat' is attached to the email.

Microsoft Outlook (Right): The email header shows the same sender and subject. The body of the email is fully visible, showing a greeting 'guten Tag' and a paragraph about electronic billing. The text reads: 'wir stellen den Kunden sämtliche Abrechnungsinformationen zunehmend nur noch elektronisch zur Verfügung. Dies erfolgt heute noch per Mail und innerhalb der kommenden 2-3 Jahre sukzessive nur noch per Portal-Login. [redacted] nschutzrechtlichen Gründen ist es uns leider untersagt, die Sammelfiles mit allen Abrechnungsdaten allen [redacted] schicken. Die Digitalisierungswelle erreicht zunehmend auch die Verwaltungen... Das Splitting von Files nimmt bei uns inkl. Nachkontrolle immer mehr Zeit in Anspruch. Aus diesem Grund ist es für uns unerlässlich, dass wir die Einzelabrechnungen inskünftig auch als einzelne pdf-Files erhalten können.'

Another winmail.dat case

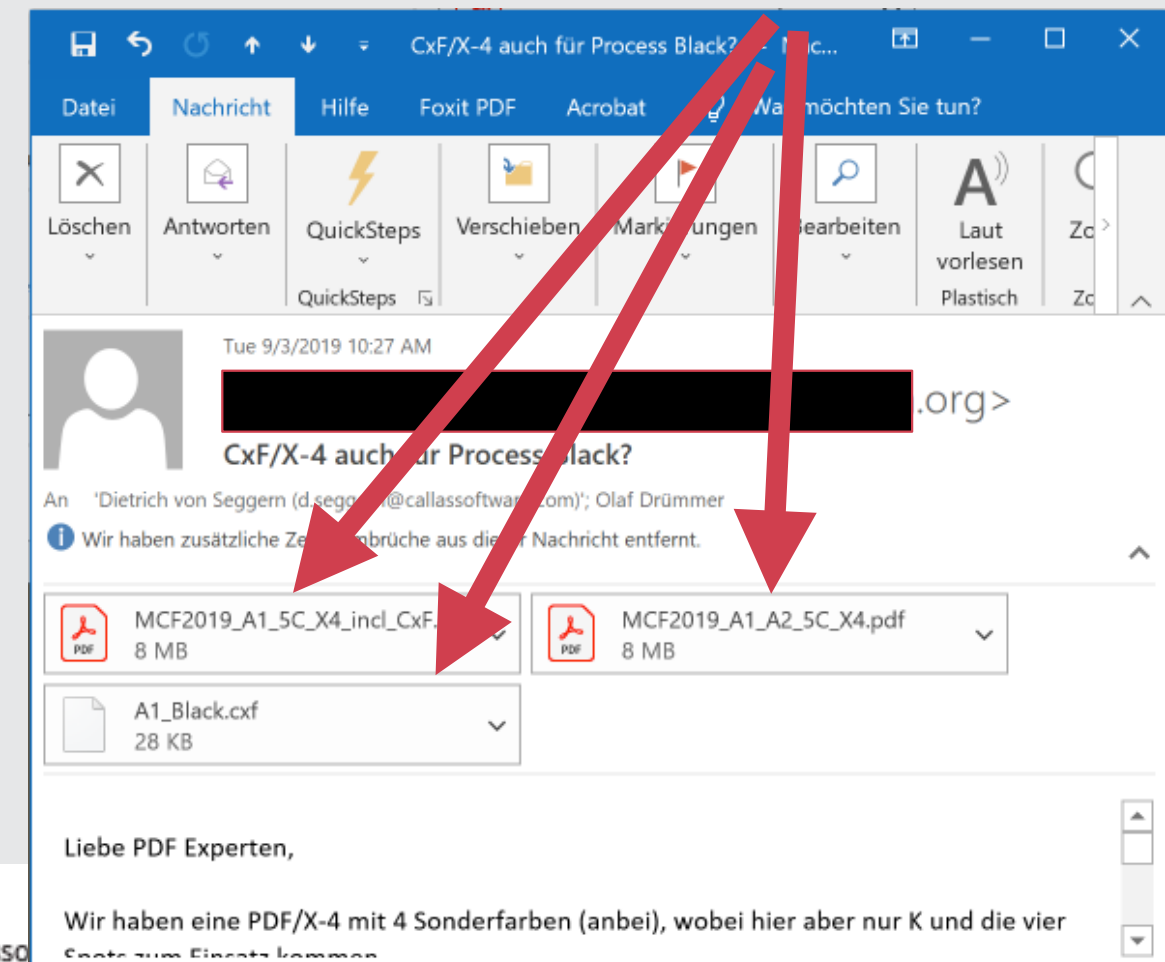
Apple mail (eml)

One “winmail.dat” attachment



Microsoft Outlook

More useful attachments



What is “winmail.dat”?



- Format specified in “Transport Neutral Encapsulation Format” (TNEF)
- “Proprietary email attachment format used by Microsoft Outlook and Microsoft Exchange Server to package special information” (Wikipedia)

How should email (incl. TNEF) and other components be represented when archived?

EA-PDF concludes



- Email archival has to include some kind of format conversion
- PDF as email archival format makes a lot of sense
 - Does not require proprietary software to be viewed
 - Established as general archiving standard for other documents



The email package file format

Email as a communication means



- Email is different from regular documents
- It is a form of communication,
a single email is usually better understood in context (mailbox)
- What file formats exist for mailboxes?

PST and MBOX



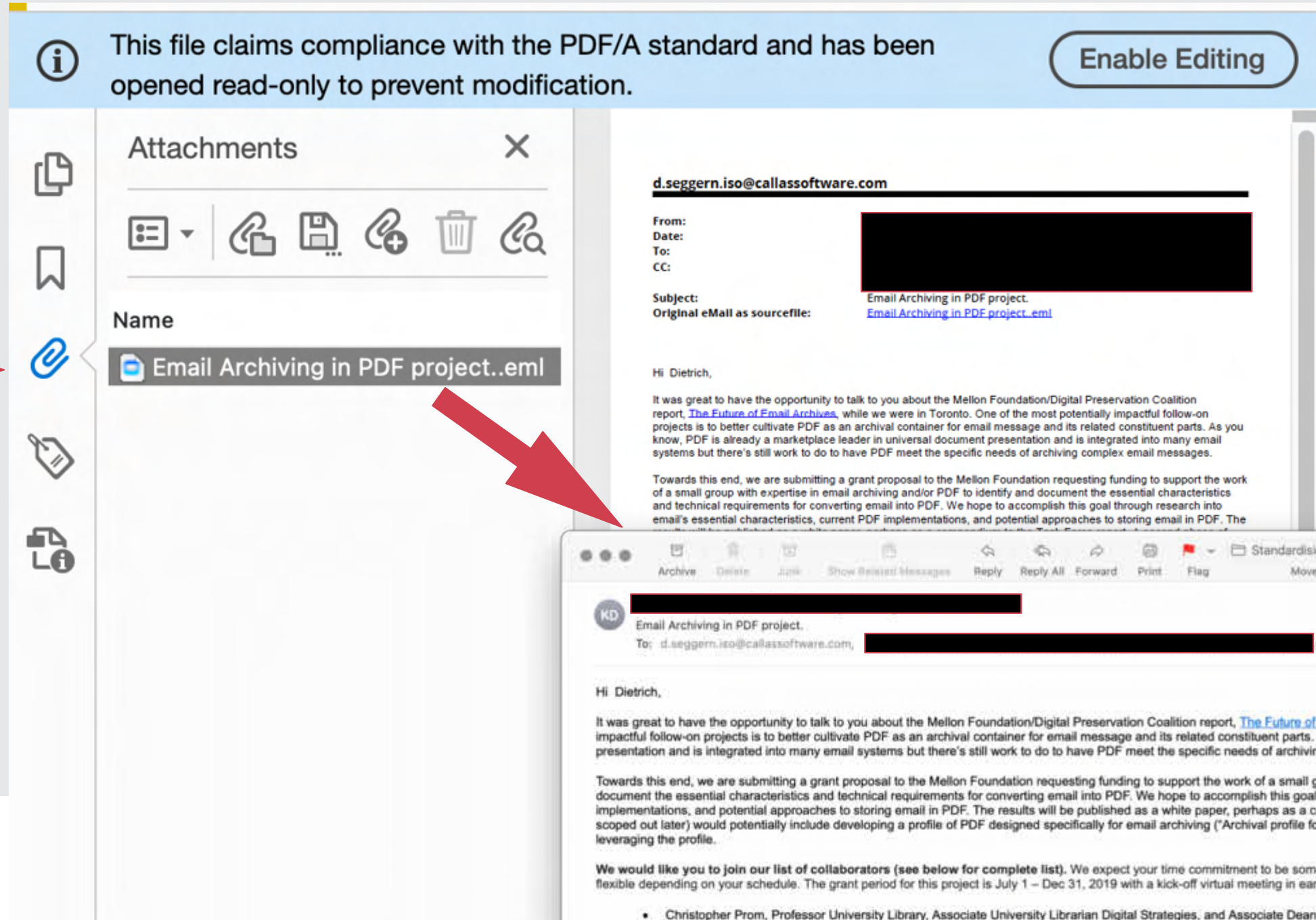
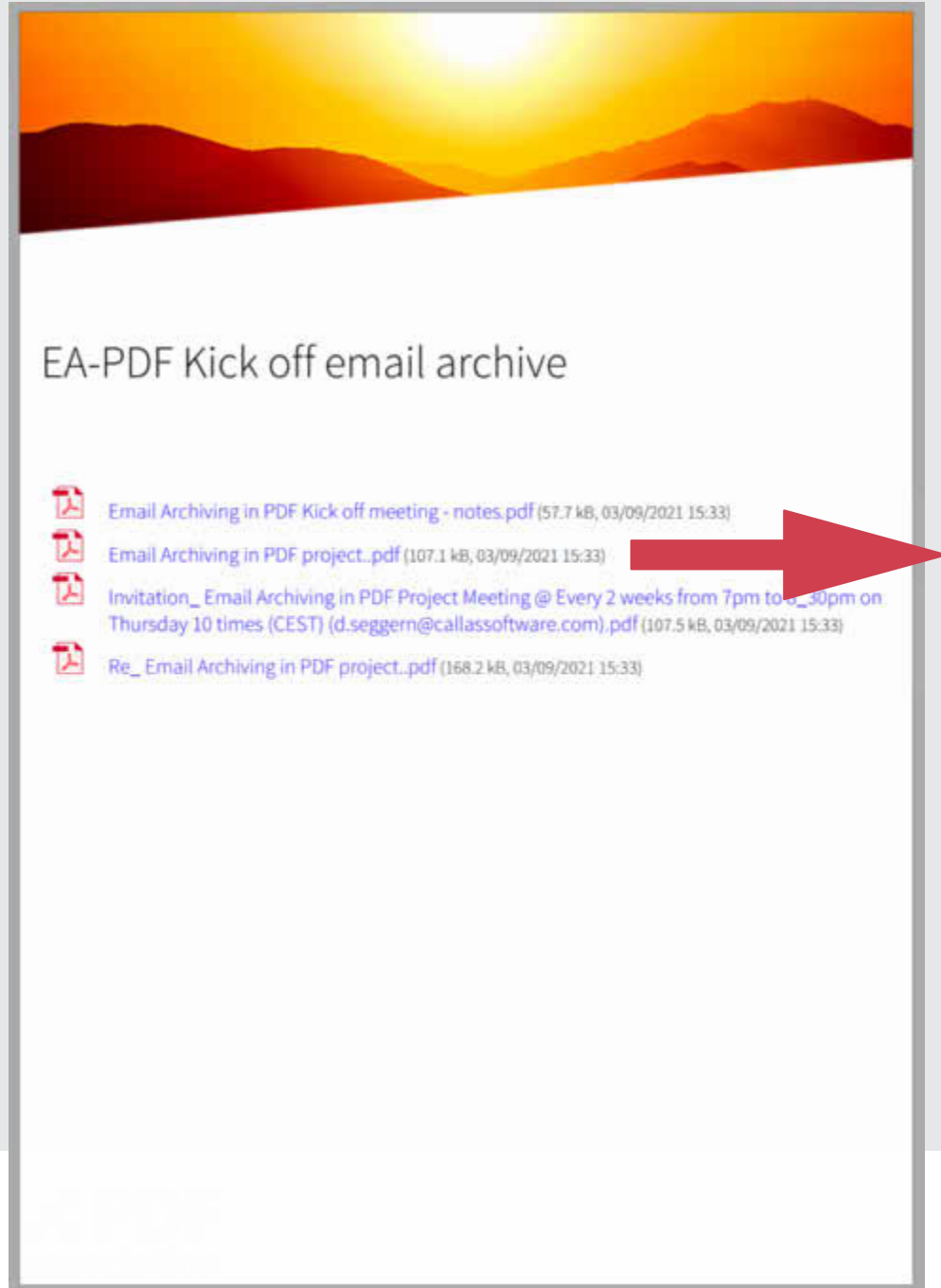
- PST - Microsoft Personal Folders File Format
 - Fully documented but proprietary Microsoft
 - Can only be opened by proprietary software
- MBOX suffix is used for for a family of related file formats, loosely standardized in RFC 41559
 - Applications can only open one MBOX format variant
- Both change when being opened

PDF in EA-PDF has two roles



- Core format that is used for individual emails
 - Plus an EML representation of the email in original format
- Container for emails if there is more than one (entire mailboxes)

How it might look like





EA-PDF functional specification for phase 2

The inner structure of an email

Email

Header

Originator

From, Sender, ReplyTo

Destination

To, CC, BCC (in sent box)

Trace

Received, Return-Path,
SPAM, etc ...

Body 3

Body 2

Body 1

MIME Part



Content-Type,
Transfer-Encoding,
Etc ...

Body

ascii or *rtf* or *html*
or data making up
an *attachment*

1) Header / Metadata



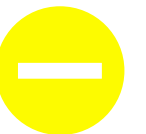
- Header field name and header field value 
- Additional descriptive metadata on EA-PDF package level according to a common schema, e.g. PREMIS 

My personal assessment

Available/Straightforward



Partly straightforward



Not available at all



To be discussed



2) Body



- All body parts

- PDF representation



- An EML representation



- PDF/A is “should”

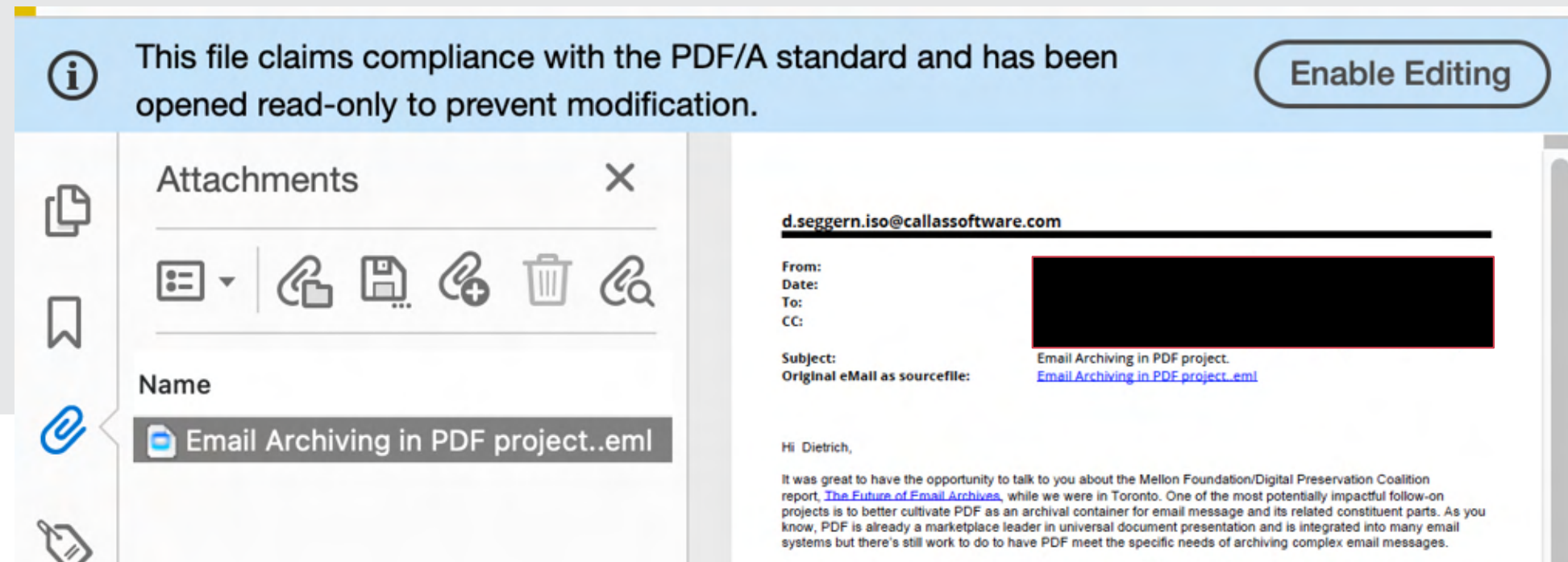


Fonts are not needed to represent a simple text email



2) Body: The core representation

- Individual emails meeting user experiences ✓
- One or more message body parts ✓
- Message body parts in rich formats, particularly HTML ✓
- Sender and recipient's email address ✓



3) Attachments







- All attachments to be included as embedded files








3) PDF as container for emails



- Extract individual emails as PDF or EML files 
- Should be digitally signed 
- The email collection should be searchable 
 - by common fields (e.g. “sender”)
 - or free text
- Sortable and filterable by common fields 
(e.g., order by date sent, or by sender)

Summary: EA-PDF's requirements



- On file format level:  
Convert emails and various email “extensions” (winmail.dat, calendar invites etc.) to PDF and EML in a useful manner
- On package level: 
Embed other files into PDF
- Retrieval:  
Searching and sorting in an EA-PDF package



Interested in next steps (phase 2)?



Join the
EA-PDF Liaison Working Group
hosted by the PDF Association

- via the member area
- via an email to info@pdfa.org

A screenshot of a web browser displaying the PDF Association's website. The browser's address bar shows the URL 'https://www.pdfa.org/community/ea-pdf-lwg/'. The page features the PDF Association logo at the top left, a navigation menu with links like 'Solution Agent', 'About Us', 'Discover', 'News', 'Events', 'Resources', 'Community', and 'Member', and a 'MEMBER AREA' button. Below the navigation, there are three highlighted sections: 'Technical resources', 'PDF Days Online 2021 Agenda', and 'Become a Member!'. The main content area is titled 'EA-PDF LWG' and contains several paragraphs of text. On the right side, there is a 'COMMUNITY CHAIR' section with two portraits: Peter Wyatt and Prof. Chris Prom, both from the University of Illinois. The bottom left corner of the page shows the PDF Association logo again.

Other questions?

Please use the [chat pod](#)
or write an email to dietrich.seggern@pdfa.org



The email to PDF/A reality show

Oct. 20 at 16:00 CEST / 10:00 ET / 07:00 PT

What of EA-PDF is already available (in a product)?



How PDF/A-2 and -4 are better than PDF/A-1

Oct. 21 at 1630 CEST / 1030 ET / 0730 PT

Why should you consider to update your archive format?