



Dietrich
von Seggern

Managing Director
callas software

Vice Chair PDF
Association



Email archival in PDF

How does that make sense?

Email archival in PDF - how does that make sense?

- What is the best way to archive email?
- Archive email as PDF?





Intro



Email (archival) today

- Essential part of today's business communication
- Business records documenting discussions, decisions, and actions
- Archival
 - Responsibilities imposed by legal environment for archival of business records
 - Business archival needs of the organizations themselves



Email is “native digital”

- Email is “native digital” and always comes with metadata
- For digital archival it should be much easier and straightforward than regular, paper based communication

BUT: In many organizations, email rarely makes its way from users’ individual accounts into records management and archival systems

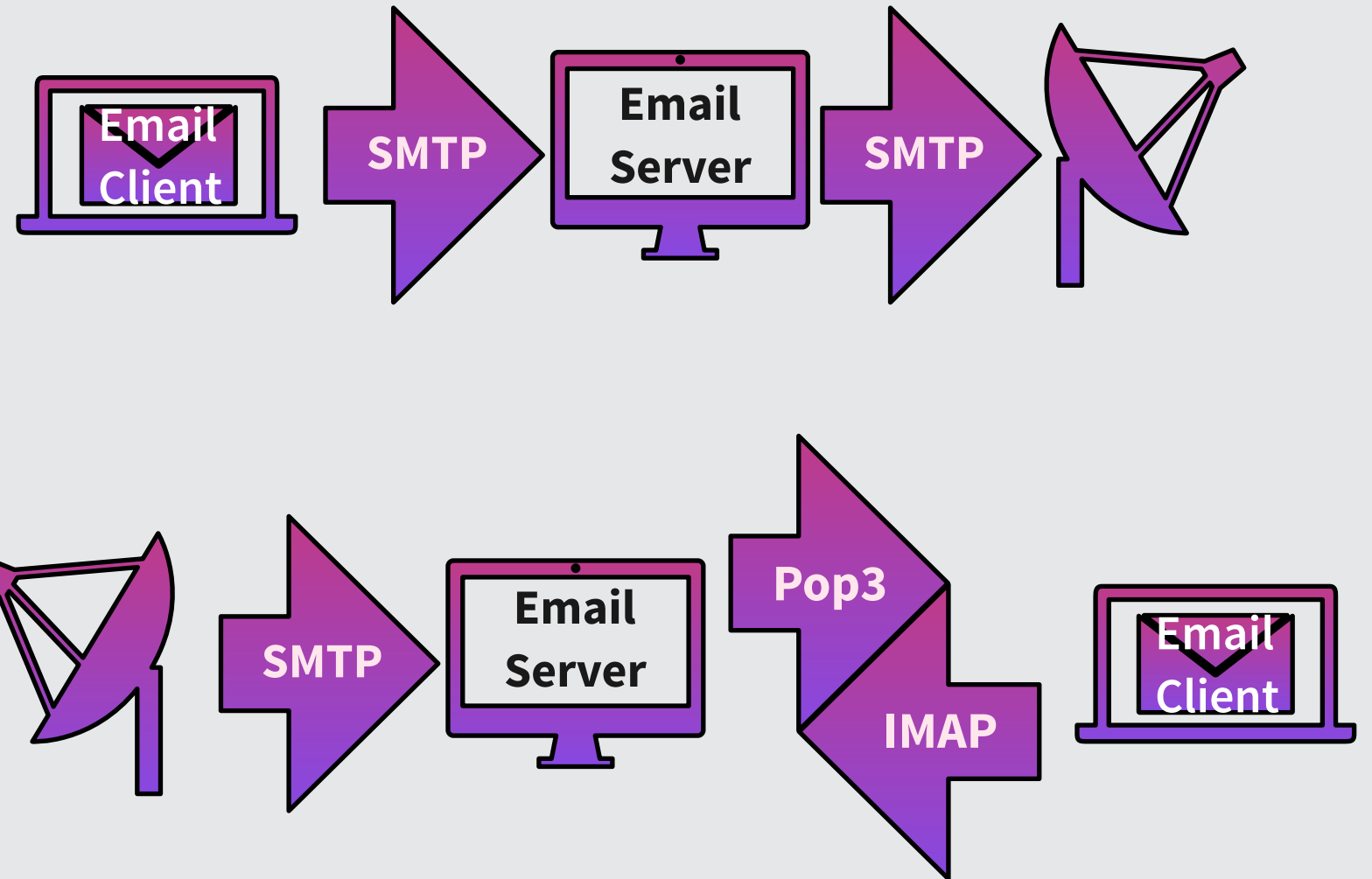
Why?



**What is the best way
to archive email?**

Email is not a file

- Email is a system of tools for composition, transport, viewing and storage
- In an ideal world we would have distributed, interoperable mailbox archives



But we can only archive files or mailboxes, so...

- ... what is the record copy?
- The email that was composed and sent?
- The copy stored on the recipient's email server system?
- A downloaded copy of it as a file?
- A copy stored locally in a PST file?


- Or does that not matter since that are all only instances of the same file?



How does email work?

SMTP (Envelope)

From
To
Received
Return Path
Data



Email file

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*



Envelop and email file are distinct

- The SMTP envelope controls the routing of the email
- Fields in the message header are for reference only (they do e.g. not contain BCC receivers)
- Typically tracing information is recorded by systems on their way, but that is not a requirement



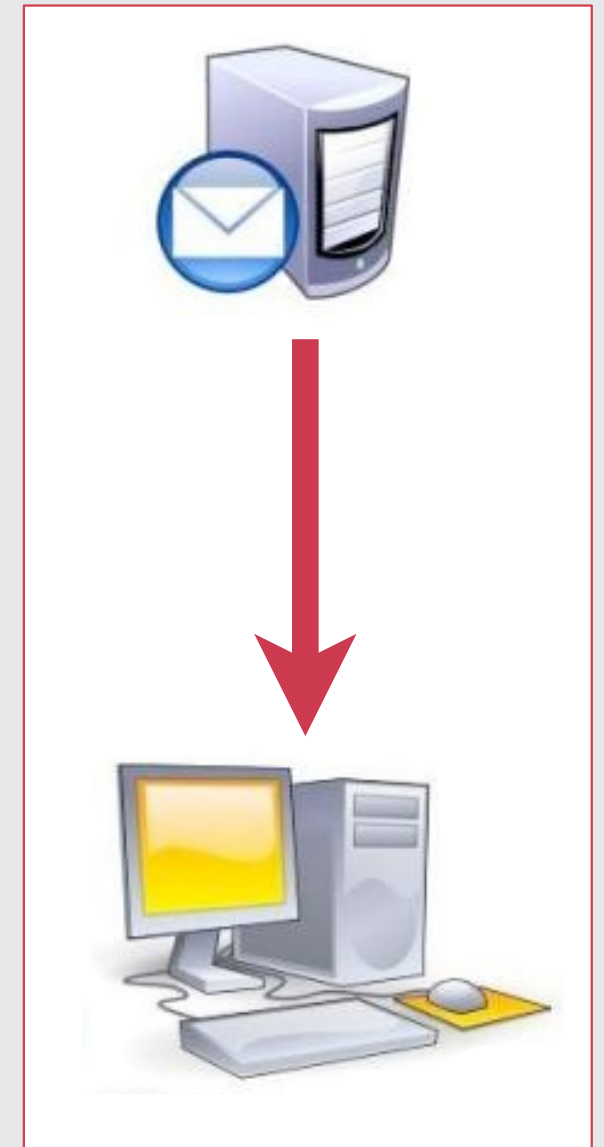
So, what is the email file format?

- The main standard for the message structure is *RFC #5322*
- Communication is standardized (information as passed) as the “Internet Message Format”
- Email servers may locally store messages in a similar format, *EML*, but that is not standardized nor a requirement



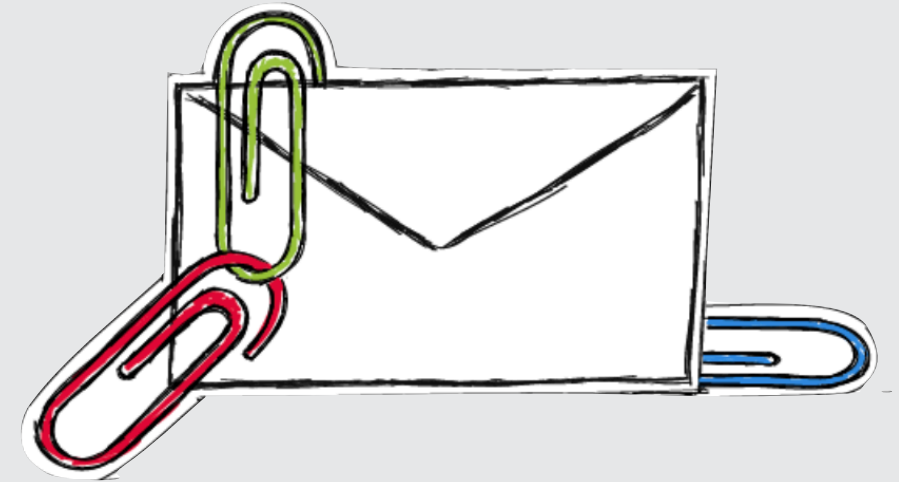
So, what is the email file format for the client?

- Email clients may leave emails on the server (IMAP) or download (POP3)
- For POP3 local formats vary, e.g. Microsoft uses a “somewhat” documented format, *MSG*
- EML is only partially documented through RFC#5322, but there is no available standards documentation



Attachments and links to external data

- Email attachments are another difficult challenge
 - Same archival questions as for any binary data:
Availability of viewer applications in the future
 - Each system saves attachments differently
-
- What about references to out-of-message content from the email body (HTML or plain text)?



Problem statement

- We can't archive the email system, only files
- It is not clear what file format has to be archived, there is no standard
- EML is a format close to what is defined in IETF RFC#5322
 - but can't always be created and
 - is not the same everywhere (not standardized)
- In addition: What is the solution
 - for attachments?
 - for formatted emails (RTF)?
 - for formatted emails in HTML with external links?



Archive email as PDF?!

What we get for email archival in PDF (or PDF/A)

■ Header

- Pendant to the letter head in classic mail
- (Mail routing takes place via SMTP)

■ Body

- Possibly a combination of content parts
 - Pure text (7 bit ASCII) *and*
 - Simple formatting via RTF, incl. enhanced type set (e.g. ä, ö, ü, œ, æ) *and*
 - Rich formatting via HTML, frequently with external links
- No guarantee for equivalence of content parts

■ Attachments

- Encoded in ASCII
- Documents
- “Archives” (Zip)
- Executables (Exe)



PDF export

- Email to PDF export is available in almost all email clients
- Migration is often incomplete and only converts the body - not header nor attachments



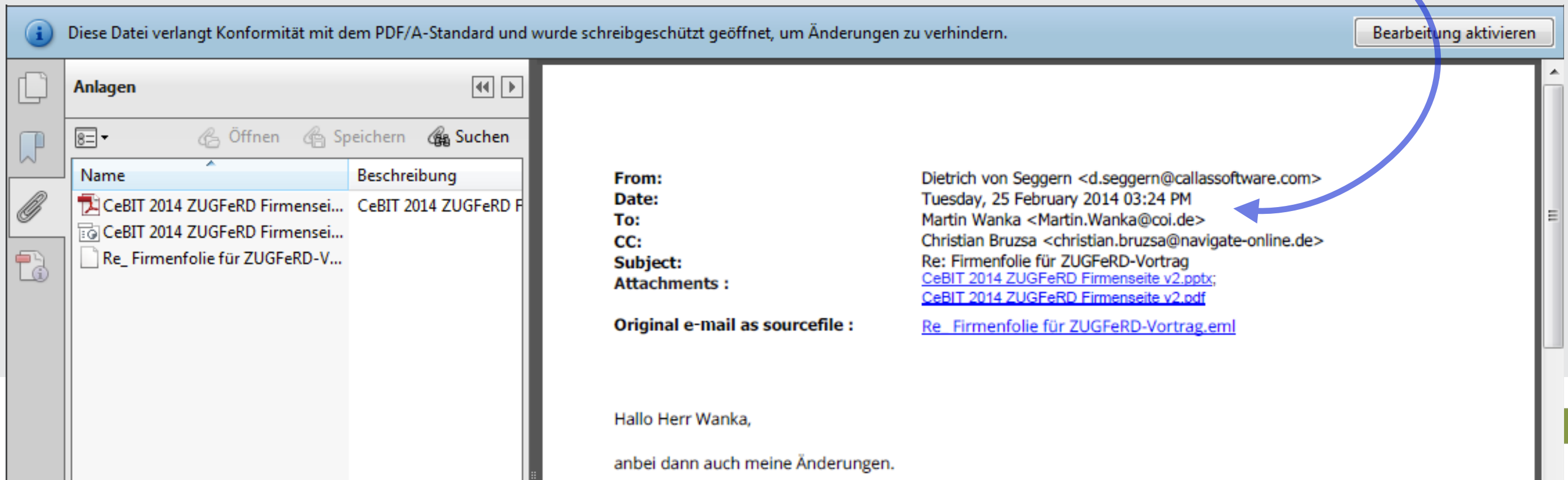
PDF export

- Email to PDF export is available in almost all email clients
 - Migration is often incomplete and only converts the body - not header nor attachments
-
- **What should an email to PDF export do?**



The email header

- Insert all header fields into the PDF's XMP metadata
- Whenever possible use predefined XMP properties or add a PDF/A Extension Schema
- Add the most important header fields also into page content (From, X-Envelope-To, Date etc.)



Diese Datei verlangt Konformität mit dem PDF/A-Standard und wurde schreibgeschützt geöffnet, um Änderungen zu verhindern. Bearbeitung aktivieren

Anlagen

Öffnen Speichern Suchen

Name	Beschreibung
CeBIT 2014 ZUGFeRD Firmensei...	CeBIT 2014 ZUGFeRD F
CeBIT 2014 ZUGFeRD Firmensei...	
Re_ Firmenfolie für ZUGFeRD-V...	

From: Dietrich von Seggern <d.seggern@callassoftware.com>
Date: Tuesday, 25 February 2014 03:24 PM
To: Martin Wanka <Martin.Wanka@coi.de>
CC: Christian Bruzsa <christian.bruzsa@navigate-online.de>
Subject: Re: Firmenfolie für ZUGFeRD-Vortrag
Attachments : [CeBIT 2014 ZUGFeRD Firmenseite v2.pptx](#);
[CeBIT 2014 ZUGFeRD Firmenseite v2.pdf](#)
Original e-mail as sourcefile : [Re_ Firmenfolie für ZUGFeRD-Vortrag.eml](#)

Hallo Herr Wanka,
anbei dann auch meine Änderungen.

The email body

- Text, RTF or HTML
 - Use that variant that has the richest content: HTML, RTF, plain text
 - At least if that is the variant that is used by the client
- For HTML bodies download and insert referenced images
- Add useful page breaks (for HTML emails)
- Embed the original email file (.eml, .msg)

Diese Datei verlangt Konformität mit dem PDF/A-Standard und wurde schreibgeschützt geöffnet, um Änderungen zu verhindern. Bearbeitung aktivieren

Anlagen

Name	Beschreibung
CeBIT 2014 ZUGFeRD Firmensei...	CeBIT 2014 ZUGFeRD F
CeBIT 2014 ZUGFeRD Firmensei...	
Re_ Firmenfolie für ZUGFeRD-V...	

From: Dietrich von Seggern <d.seggern@callassoftware.com>
Date: Tuesday, 25 February 2014 03:24 PM
To: Martin Wanka <Martin.Wanka@coi.de>
CC: Christian Bruzsa <christian.bruzsa@navigate-online.de>
Subject: Re: Firmenfolie für ZUGFeRD-Vortrag
Attachments : [CeBIT 2014 ZUGFeRD Firmenseite v2.pptx](#);
[CeBIT 2014 ZUGFeRD Firmenseite v2.pdf](#)
Original e-mail as sourcefile : [Re_ Firmenfolie für ZUGFeRD-Vortrag.eml](#)

A blue arrow points from the text "Embed the original email file (.eml, .msg)" in the list above to the "Original e-mail as sourcefile" link in the screenshot.

Email attachments

- Use PDF/A-2 or PDF/A-3
- Either embed original and PDF/A - or define individual format rules (whitelist, blacklist)
- Unzip ZIPs
 - Store folder structure in the PDF
- Create file links in the PDF file



What about file sizes?

- Original file formats
 - File transmission protocol uses ASCII 7 Bit (for backwards compatibility)
 - Binary files increase file size significantly (about 75%)
 - The size of the saved file is application dependent
- PDF
 - Makes extensive use of compression
 - PDF/A requires more information to be stored for completeness, e.g. embedded fonts or device independent color (ICC profiles)



Some personal statistics

- “Regular” text email without attachments
 - PDF/A is ca. ~2 times bigger
- Text email with binary attachment (office file)
 - PDF/A has almost the same size
- HTML email
 - PDF/A is significantly bigger, but only due to referenced images



Retrieval

- Not sufficiently addressed
- Relationships between email instances in different mailboxes
- Better search features - e.g. show me all threads:
 - that spread over more than 2 months
 - in 2009 or 2010 and
 - where colleague A and external B were involved
- *It is even more important to archive all available information*



Problem statement - What can be addressed with PDF?

- We can't archive the email system, only files
- ✓ It is not clear what file format has to be archived, there is no standard
- ✓ EML is a format close to what is defined in IETF RFC#5322
 - ✓ but can't always be created and
 - ✓ is not the same everywhere (not standardized)
- ✓ In addition: What is the solution
 - ✓ for attachments?
 - ✓ for formatted emails (RTF)?
 - ✓ for formatted emails in HTML with external links?



Conclusion

- PDF/A can be used as archive format for email
 - Emails can be converted into PDF/A with the “look and feel” of the email client
 - Attachments can be converted to PDF/A as well and embedded into the PDF/A “body”
 - PDF/A-3 allows for embedding the original attachment as well
 - All header information should be added to the PDF metadata
- PDF/A attachments allow for system independence and completeness, they therefore can fully avoid complex migration concepts and projects
- This approach stores as much information as currently possible and allows for best possible retrieval in the future





Questions?

Comments are welcomed.



Thank you!

Dietrich von Seggern

Managing Director callas software

Vice Chair PDF Association

We appreciate your participation.

