# Requirements for Archiving Email using PDF

**Draft Report for Comment & Review**

**April 29, 2020**

## Table of Contents


Executive Summary

Introduction

- Audience
- Purpose
- Working Group Members
- Process

Objectives and Background

- Why Email to PDF?
- Other approaches
- Technical background
- PDF features of special interest for EA-PDF
- Nomenclature
- Terms and definitions

Functional requirements

- 1 Open standards
- 2 Capturing Email
  - 2.1 Scope of emails to be captured
  - 2.2 Capturing Emails as embedded files
  - 2.3 Email header fields
  - 2.4 Body parts and attachments
  - 2.5 Additional message metadata
  - 2.6 Linked content

2.8 Considerations for legacy PDF readers

3 Describing Email

3.1 Recording explicit exclusions

3.2 Describing the scope of an EA-PDF archive

3.3 Provenance

3.4 Additional metadata

4 Representing email

4.1 Core representation of individual email messages

4.2 Collections of email messages

4.3 Describing the archive within the core representation

5 Functional requirements for EA-PDF readers

5.1 Display and search

5.2 Email extraction

5.3 Content extraction

5.4 Metadata extraction

5.5 Attachment representation and extraction

Appendix A - Problems with existing email formats

Common message storage formats

PST - The Microsoft Personal Folders File Format (PST)

MBOX

EML (Electronic Mail Format)

Appendix B - Privacy and ethical concerns

Appendix C - Metadata Options

# Executive Summary

Will be added after comments have been received and incorporated.

# Introduction

This document establishes fundamental requirements for archiving email and sets out an approach to considering ISO 32000 Portable Document Format (PDF) technology as a model for capturing email for long-term archival purposes using open, ISO-standardized technologies. Its development was supported by a grant from the Andrew W. Mellon Foundation to the University of Illinois at Urbana-Champaign. A public version of the project rationale and description is available at: https://emailarchivestaskforce.org.

## Audience

verify authenticity (e.g., digital signatures), all in machine-readable form

We expect that the following groups of people will find this requirements document to be of interest:

- Archivists, digital preservation professionals, records managers, and curators considering advanced email archiving stewardship and technology.
- Government, discovery, and legal records-management practitioners.
- IT professionals administering enterprise communications systems.
- Implementers of PDF and email technology who are interested in understanding the needs of digital preservation professionals.

## Purpose

These recommendations are intended as a framework within which interested people from the archives, digital preservation, and PDF communities can collaborate. We hope that a technically detailed specification and implementation reference model will be developed, using the requirements outlined here as a starting point. Ultimately, we seek to describe how advanced PDF-based capture, preservation, rendering, and distribution options for email correspondence can deliver long-term value to organizations, communities, and members of the public.

## Working Group Members

The following people contributed to the development of this document:

- Christopher Prom, Principal Investigator. Associate Dean for Digital Strategies, University of Illinois at Urbana-Champaign
- Joel Simpson, Project Consultant and Executive Committee Member, Artefactual Systems.
- Kevin De Vorsey, Executive Committee Member. Senior Electronic Records Policy Analyst, National Archives and Records Administration
- Kate Murray, Executive Committee Member. Digital Projects Coordinator, Library of Congress
- Christopher (Cal) Lee. Professor, School of Information and Library Science, University of North Carolina at Chapel Hill
- Steve Levenson, ISO TC 171 SC2 WG5 Convenor
- Camille Tyndall Watson, Digital Services Section Head, State Archives of North Carolina
- Jamie Patrick Burns, State Archives of North Carolina

- Tricia Patterson, Digital Preservation Analyst, Harvard Library
- Lynda Schmitz Fuhrig, Digital Archivist, Smithsonian Institution Archives
- Stephen Abrams, Head of Digital Preservation Services, Harvard Library
- Dietrich von Seggern, callas software, GmbH
- Duff Johnson, Chief Executive Officer, PDF Association
- Matthew Hardy, Sr. Engineering Manager, Document Cloud, Adobe

## Process

The working group held a two-day in-person meeting at the Library of Congress, on November 5 & 6 2019, to review project goals, email and PDF functionalities, to review a very early draft of this document, and to set a drafting and editing schedule. For the next 4 months, we met online biweekly via Zoom to discuss open questions and make decisions, leading to the recommendations in this draft. Specific feedback can be provided via two methods:

- Commenting on this document using Google Docs comment features: This is recommended for feedback on specific points, requests for clarification etc.
- Sending an email to the group via this web form: https://bit.ly/2PRk5cL. This is recommended for longer, more substantive comments.

All feedback will be reviewed by the group and incorporated during April before the final release of this report by June 30th, 2020.

# Objectives and Background

This document defines the functional and technical principles to be elucidated in a forthcoming (Phase 2) industry specification for “EA-PDF” (Email archiving in PDF).

NOTE: ISO standardization is available, but is not contemplated or required for EA-PDF.

By developing these requirements, we set the stage for software developers to create email-to-PDF writers that retain the core metadata, content, attributes, and context that contribute to the digital object’s integrity and authenticity while also providing a standardized facility for capturing provenance metadata.

Current email-to-PDF pathways lack this ability. Today, printing an email message to PDF produces an incomplete version of the message. Most header information is ignored, much as if you printed the message on paper. By providing an email-to-PDF migration pathway that considers archival needs, we combine the possibility to both satisfy archival requirements (by writing detailed header information to metadata fields in the PDF file), and to provide a dissemination pathway independent of email software.

While emails can be exported, stored and preserved in something approaching their native formats (for example as PST, MBOX, or EML files), those files are typically only rendered and viewed with email software. Many people will not be comfortable importing others’ archived email into their own email client, for security and other reasons. Archiving email to PDF provides a ubiquitous, simpler, and more secure way to access and view archived messages.

This document defines the core archival requirements for capturing email using PDF technology. While PDF presents many opportunities to support dissemination, there is no consensus on what “core”

dissemination requirements should be. We don't set out hard dissemination requirements, but do provide some implementation guidance. We recognize that use cases differ.

## Why Email to PDF?

One may ask: where did this document come from and why is it needed? After all, aren't there already capture and migration pathways for email? And why should PDF be considered as a potential target format for archival-quality, preservation-enabled emails? These are good questions, and responses might be grouped under two general headings:

**PDF addresses gaps and risks that are inherent to current email formats and migration pathways.**

- PDF includes rich data structures to accommodate the diversity of email content and metadata. Completely self-contained PDF facilitates the capture of text and graphical content for archival purposes. It includes extensive capabilities supporting renderings (e.g. of email content), arbitrary files (e.g. email attachments), source data (e.g. IMF), metadata (e.g. header fields), and data to verify authenticity (e.g. digital signatures), all in machine-readable form with full capture of semantics.

- PDF provides a vehicle for capture of provenance metadata as part of the act of archiving a mailbox (or server, client, folder, message, etc).

- Email-to-PDF provides a migration pathway to a dissemination packet for individual or aggregated email messages. It would preserve many of the essential attributes of the message, including header metadata, in an easily distributable format that can be opened on any device that includes a basic PDF reader.

**Email-to-PDF migration leverages existing standards and a broad and diverse vendor community.**

- There are many use cases for preserving, searching, and reusing email from commodity services.

- PDF allows the ability to integrate interoperable email preservation tools into existing, widely used tools such as email servers and clients.

- As the *Future of Email Archives Report* notes, "[E]mail archiving is still an emerging practice."[1] Email as PDF could be ingested, stored, preserved, and disseminated from established, widely implemented repository systems that are already in use in government, academic, public, and corporate archives and libraries.

- Since the PDF format is so extensible and widely implemented, a common understanding of best-practices for archiving email with PDF would facilitate development of email specific viewers that can provide browsing and searching functions similar to those that exist within email client applications.

[1] Task Force on Technical Approaches for Email Archives. "The Future of Email Archives." New York: Council on Library and Information Resources, August 2018. https://www.clir.org/pubs/reports/pub175, 1.

In short, this work seeks to leverage existing technologies to allow individuals and institutions a pathway to migrate email into the most widely used and implemented format for the distribution of text documents.

## Other approaches

Archiving email as PDF does not preclude other approaches, including emulation of email systems or retention of messages in email-specific formats. PDF may complement other archiving strategies instead of replacing them. For those that do choose PDF as an archiving option for email, a standardized application of PDF technology can serve as a stable and structured means of bundling extractable email source data, universally usable archival-quality renderings and provenance metadata.

Some institutions will choose to preserve and represent email within platforms that use email-specific formats such as MBOX, EML or PST. Others may emulate old email environments or use other formats or XML schemas. These approaches require a relatively high level of technical development or support, possibly including the development of parallel discovery and access environments. Archives, libraries, and other memory institutions have experimented with these approaches, but have not widely implemented them as production services. For the many archives that are simply storing format specific email archives as unprocessed holdings, email-to-PDF offers a relatively straightforward migration pathway with demonstrated downstream usability. For example, it provides a compelling option for government and university archives that seek to disseminate large volumes of email. It can be rendered easily using the PDF readers that are built into most web browsers and operating systems; The State Library of Virginia chose PDF to distribute emails from the Virginia governor's office partly for this reason.[2]

The framework offered in this document, therefore, provides a pathway to help people do something they are already doing and may need to do given a particular set of institutional factors: convert email to PDF and to do it in a way that preserves the essential provenance metadata that allows us to say that the messages are authentic and complete.

## Technical background

Introduced in 1993 by Adobe, PDF (Portable Document Format) is a flexible multi-platform digital document format adopted worldwide. Creating PDF files is as easy as printing; viewing PDFs has always been free. As a replacement for distribution of paper documents, PDF is a proven, reliable solution.

Today, PDF is an open, standards-based technology that may be implemented by any capable developer. The technology is supported by a broad ecosystem of vendors around the world.

PDF became an ISO standard (ISO 32000) in 2008, joining PDF/A (ISO 19005), the archival subset of PDF designed for long-term preservation. PDF 2.0 (ISO 32000-2) was published in 2017.

ISO standardized and industry-supported PDF technology development is rooted in the not-for-profit and vendor-neutral PDF Association.

[2] State Library of Virginia, "Virginia Memory: Collections: Kaine: Look Under the Hood," 2016, http://www.virginiamemory.com/collections/kaine/under-the-hood.

## PDF features of special interest for EA-PDF

ISO 32000 technology includes a variety of features making it well-suited to archiving email, in particular:

- Support for full-text search
- Support for document and object-level metadata
- Support for embedded files, including email source data (e.g. IMF)
- Support for redaction, annotation, and linking functionality
- Semantic structures for rich reuse and extraction of content
- Support for authentication (digital signatures)
- Support for packaging provenance metadata (e.g EA-PDF archive creation dates)
- Support for multi-gigabyte files and millions of pages
- The PDF/A subset targeting archival needs
- The PDF/UA subset targeting accessibility needs

## Nomenclature

The following terms are normative when used herein; that is, they have specific meanings. Reading the terms for their defined meanings is essential to understanding this document.

- “shall” / “shall not” = required / prohibited
- “should” / “should not” = strongly recommended / strongly disfavored
- “may” = permitted

## Terms and definitions

We use email terms (such as header field, body part) as defined in the Internet Message Format standard (RFC 5322; https://tools.ietf.org/html/rfc5322) and the Multipurpose Internet Mail Extensions (MIME) family of standards (RFC 2045; https://tools.ietf.org/html/rfc2045).

**core representation:** email data as captured to PDF page content in an EA (Email Archiving)-PDF file

**EA-PDF:** PDF file created according to the provisions of this document

**EA-PDF writer**: software that processes input email content and writes an EA-PDF file

**EA-PDF processor:** software that reads, updates, or otherwise processes an EA-PDF file

**EA-PDF reader:** interactive viewing software that reads EA-PDF files

**email archive:** one or more EA-PDF files containing one or more archived emails or email collections

**email archive creator:** the entity (user or software) operating the EA-PDF writer (and thus, implementing policies)

**legacy PDF processor:** software that writes, reads, updates, or otherwise processes a PDF file which conforms to ISO 32000, but is unaware of (and thus, incapable of conformance with) this document

**legacy PDF reader:** interactive viewing software for PDF files that is unaware of EA-PDF

# Functional requirements

## 1 Open standards

Open (non-proprietary) standards reduce barriers to full interoperability of data and metadata (whether by humans or machines), reducing risks to long-term preservation. Widely adopted standards are supported by a greater range of tools and technology. Finally, well-written standards make it easier to develop compliant tools and technology to improve quality and reliability.

Conceptually, EA-PDF is a PDF superset specification within the context of PDF 2.0 (ISO 32000-2:2017), the presumed technical basis for implementing full-featured EA-PDF archives using PDF technology. The following standards will be indispensable to developers of EA-PDF processors:

- ISO 32000-2 (PDF 2.0)
- ISO 19005-4 (PDF/A-4), to be published in late 2020
- ISO 14289-2 (PDF/UA-2), to be published in late 2020 or early 2021
- ISO 16684-2:2014 (XMP)
- PDF Declarations, The PDF Association, https://www.pdfa.org/resource/pdf-declarations/

**1.1 An EA-PDF archive shall conform to the ISO 32000-2 (PDF 2.0) standard**

**Rationale**

- The PDF 2.0 standard is the latest of the general-purpose PDF standards; and includes several features leveraged by EA-PDF .
- Although PDF 2.0 adoption is not yet widespread, EA-PDF writers will create files that provide substantial EA-PDF functionality to users with processors that are unaware of both PDF 2.0 and EA-PDF. See the requirements in Section 4 of this document for more details.

**1.2 An EA-PDF archive should conform to the ISO 19005-4 (PDF/A-4) standard.**

**Rationale**

- The PDF/A standard exists to meet the needs of archiving institutions and incorporates requirements to ensure archival quality files.

  *Note on capture to PDF (but not PDF/A)*

  PDF/A requires embedded fonts to ensure accurate and consistent rendering irrespective of platform. Plain text email doesn't insist on particular fonts, and so some users may feel that PDF/A conformance is not appropriate when capturing such email (and/or that PDF/A's benefits are outweighed in a given case by the file size cost implied by embedding fonts.

**1.3 An EA-PDF Archive should conform to the ISO 14289-2 (PDF/UA-2) standard.**

**Rationale**

- The PDF/UA standard exists to ensure accessibility of PDF files. It incorporates a wide range of features to ensure accessibility by the widest possible set of users.

- Some accessibility features are hard or impossible to implement at the point email is captured; features like providing alternative text for images in an email may not be possible without input from the original author (who may not be available).

**1.4 An EA-PDF Archive shall consist of well-structured metadata that conforms to specified open standards.** See Appendix C: Metadata for a discussion of potential standards that can be used.

**Rationale**

- Capturing metadata provides crucial information for determining authenticity and understanding the full context of an archive.
- Using standard vocabularies and schemas that are well defined and documented helps users to reliably interpret and understand the meaning of the metadata.
- Standard schemas also improve interoperability between systems, which can significantly improve discovery, search and access.

# 2 Capturing Email

## 2.1 Scope of emails to be captured

**2.1.1 EA-PDF writers should, by default, capture all email messages in the account, volume, or file to be archived, including any messages held in folders or tagged by special labels, such as sent items, deleted items, drafts etc., leaving it to email archive creators to specify any email messages to be excluded.**

> NOTE: EA-PDF archives can contain one or more email messages, as the scope or purpose of a given email archive is entirely up to the creator or institutional context of creation. Valid email archiving use cases range from individual messages to entire mailbox files/collections or server instantiations.

**Rationale**

- Although archive creators may have good reason to exclude certain folders or messages, EA-PDF writers should not assume content should be excluded.

- Best practice in email archiving includes capture of all email available at the moment of export, and to make appraisal or selection decisions at a later date.

**2.1.2 EA-PDF writers should allow EA-PDF creators to opt to retain email content that fails virus-scanning checks.**

**Rationale**

- Although common practice is to delete malicious or suspected content, some institutions' policies may require retention of all data without exception.

## 2.2 Capturing Emails as embedded files

**2.2.1 EA-PDF writers shall include each individual email as an embedded file using the Electronic Mail Format.** (See Appendix A for a description of this format)

**Rationale**

- A wide variety of applications and systems (from generic document management systems to specialized email processing tools) can parse, render, or interpret emails based on the ubiquitous email standards.
- The Electronic Mail Format is essentially the practice of writing out messages that conform to the Internet Message Format (IMF) standard and the Multipurpose Internet Mail Extensions (MIME) standards to a text file with an extension of .eml. It is the closest approximation of the "original" format of an email and has widespread support in many email applications.

## 2.3 Email header fields

**2.3.1 All header fields, including both the *header field name* and *header field body* shall be captured unless institutional policy explicitly requires that they be excluded.**

**Rationale**

- Header fields are often essential to understanding the authenticity and context of the record; some header fields are essential to understanding the structure of the record (because they reference other parts of the email, in particular body parts).
- Even custom or optional header fields can be widely used and adopted; email systems with wide adoption (and universal adoption in the case of particular organisations) such as Microsoft Exchange utilize "optional" header fields that are not defined by any RFC - but may well be considered "significant" and well understood in a particular institutional context.
- All of the existing standard email export formats (.mbox, .pst, .msg, etc.) retain all header fields.
- There may be valid reasons to exclude particular header fields (header field name and/or header field body) such as privacy, meeting security or classified information policies, etc. It's beyond the scope of this document to define valid reasons for such exclusions.
- Header fields are usually short, simple text (thus no strong argument for exclusion due to size or storage).

**2.3.2 All captured header fields shall be individually tagged in appropriate metadata to maximize support for diverse downstream uses.**

**Rationale**

- Granular capture of header fields enriches display options, search, aggregation, and other downstream processing.

## 2.4 Body parts and attachments

**2.4.1 All *body parts* and *attachments* shall be captured unless the email archive creator provides an explicit reason to exclude them.**

**Rationale**

- While some body parts provide the same text in different formats (e.g., a plain text version and an HTML version), retaining all versions provides future users with potentially useful evidence and allows those users to determine which version has most relevance.
- Attachments are commonly as important as the message, or more so.
- There may be valid reasons to exclude particular body parts or attachments such as protecting privacy, meeting security, or classified information policies, etc. Such considerations are out of scope for this document.

**2.4.2 All non-excluded email attachments shall be included as embedded files.**

**Rationale**

- Since the format of attachments is inherently arbitrary, email users do not expect all attachments to be directly usable or renderable in an email application.
- Users expect to be able to extract attachments for use by other applications.

## 2.5 Additional message metadata

**2.5.1 When available for capture, additional message metadata (i.e., metadata not included in header fields) should be captured with full granularity. Stored data should reference any applicable schemas the metadata may conform to (for example, IMAP fields).**

**Rationale**

- Many email applications store additional metadata beyond what is included in email header fields, which may be highly useful to future users. Examples include flags indicating whether a message was read or not, the importance or urgency of a message, which folder a particular email was stored in, or descriptive labels.

- Some additional metadata may conform to defined standards, such as IMAP fields. Referencing the IMAP standard used (e.g. the flag attribute) increases the chances that future users can understand and use this metadata effectively.

## 2.6 Linked content

**2.6.1 EA-PDF writers should allow creators to capture *linked content* that is intended for inline presentation at the time of archive creation.**

**Rationale**

- Emails commonly contain links to external content that is presented inline to email users. Most commonly this includes links to images that are presented within the email message. While users may perceive the image to be “part of the email,” in fact, the image is being retrieved by email clients at the time of presentation.

- The ability to retrieve linked content degrades over time (“link rot”). The earlier linked content is captured, the greater the chance that the intended content is captured accurately.

**2.6.2 EA-PDF writers may allow creators to capture any *links to external resources* at the time of archive creation.**

**Rationale**

- Emails commonly contain links (such as <a href> links used in HTML) to external resources that may have curatorial value. Links are intended for the email user to decide whether they would like to retrieve those resources.

- While links to external resources are less likely to be considered “part of the email,” they may still be essential to the broader purpose of the email, and have considerable value to future researchers, historians, etc.

### 2.8 Considerations for legacy PDF readers

**2.8.1. EA-PDF writers shall be capable of producing EA-PDF files intended for legacy PDF readers. Fallback functionality and implications shall be explicitly defined for EA-PDF creators.**

**Rationale**

- EA-PDF readers may not be available to all users.

- Maintaining equivalent or acceptable functionality in legacy PDF readers may restrict EA-PDF archive creation options (e.g., an EA-PDF reader may include specialized search software processing email metadata in EA-PDF files, whereas a legacy PDF reader may only include simple text search facilities within the core representation).

## 3 Describing Email

### 3.1 Recording explicit exclusions

**3.1.1 EA-PDF writers should document any explicit exclusions of email content or metadata, including what specifically was excluded (a header field, part of body part, an attachment, etc.) and the reason for the exclusion (privacy, virus, etc.) using PREMIS or other similar suitable model/schema in extractable form.**

**Rationale**

- Documenting what content or metadata was excluded and why allows future users to gain a much more accurate and complete understanding of the archive.

## 3.2 Describing the scope of an EA-PDF archive

**3.2.1 EA-PDF writers should include metadata describing the scope of an archive, including any curatorial selection criteria (e.g. date range, emails from particular folders, topical theme, strategic priority, etc.).**

**Rationale**

- Understanding the choices made in creating a particular archive provides critical context for future users.
- Describing the scope of the archive will help future users understand how “complete” or comprehensive the archive is and provide some evidence of emails or metadata that is not in scope.

## 3.3 Provenance

**3.3.1 EA-PDF writers shall include archive creation date and author (user, software, institution) using PREMIS or other similar suitable model or schema. Minimum required fields include:**

- Archive creation date (the date the EA-PDF Archive was created)
- Archive creation software (the software responsible for generating the EA-PDF Archive)
- Archive source (the file(s), client software, and/or servers used as a source)

**Rationale**

- Capturing metadata about provenance is a fundamental archival practice; understanding the source and context of records is critical to understand the records themselves.
- This information is easily obtained and captured by any software generating a digital object.
- This is one of the major gaps in current email formats; many formats have very little information about how the digital object was created (when, by what user, using what software, using what criteria, etc.).

**3.3.2 As available, EA-PDF archives should include details of the original creation and subsequent processing of the email using PREMIS or other similar suitable model/schema in extractable form. If captured, such data shall be captured with full granularity including, at a minimum:**

- Account holder
- Email domain
- Institution that hosted that domain
- The person associated with the email account

**Rationale**

- This information provides valuable context for future users to understand the source and nature of the content included in the EA-PDF Archive.

- This is presented as a “should” requirement because it may not be easy to obtain; account level information is not covered by the core email standards and will vary from one email system to another; alternatively it may require user input

**3.3.3 EA-PDF files should be digitally signed by the authoring institution and user.**

**Rationale**

- Digital signatures are a mechanism for verifying the authenticity and integrity of a digital object.

### 3.4 Additional metadata

**3.4.1 EA-PDF writers shall provide facilities for users to add additional metadata to the EA-PDF archives they create, including metadata that describes:**

- Entire emails
- Portions of text within the message of a particular email
- Other metadata (e.g., information specific to a server rather than an account)
- Attachments

**Rationale**

- There are many reasons for marking up content within an email archive, such as identifying sensitive information (to prevent inappropriate disclosure) or identifying topics or other descriptive information that may help later users.
- It is not in the scope of this document to suggest particular practices or procedures (whether, how, or when these activities should be done). Though recognising that this is a common practice in many institutions, providing a standard framework will improve the chances that future users will be able to make use of this metadata.
- Providing a standard mechanism for marking up the content of EA-PDF Archives will improve interoperability and future use; for example, to make use of redaction features in existing PDF readers.

## 4 Representing email

It is to be expected that the experience of EA-PDF files will differ between EA-PDF readers and legacy PDF readers. The core representation specified in this clause provides the basis for a common experience of the archive.

The EA-PDF core representation provides an easy access dissemination mechanism due to the widespread availability of PDF readers. While EA-PDF allows users to extract messages in a format suitable for email software, the core representation allows users to access the archive’s content directly. As a result, institutions preserving email do not necessarily need to maintain additional software for access or dissemination.

Providing the essential elements of emails in the core representation supports backward compatibility with legacy PDF readers.

EA-PDF does not specify the order, prominence, or format of the core representation - these can vary from one email application to another and thus can be determined by implementers of the EA-PDF standard.

## 4.1 Core representation of individual email messages

**4.1.1 The core representation should provide an experience of individual emails similar to that which users typically see in common email applications.**

**Rationale**

- Visual cues associated with displayed email help users understand the content's context as experienced in the email viewer.

**4.1.2 The core representation shall display one or more parts of a multi-part message.**

**Rationale**

- The main message of emails are often provided using different formats (most commonly one in text format and one in HTML).
- A multi-part message includes message content provided in more than one format; (e.g., one body part with `Content-Type: text/plain` and another version of the same body part with `Content-Type: text/html`)
- While text formats are readable for human users, added formatting - such as bolded fonts or embedded tables - can represent significant information and are more easily readable and offer a more intuitive usability experience.
- Curators wanting to avoid the additional space one copy of each format would occupy should have the ability to opt for only one.

**4.1.3 Body parts in rich formats, particularly HTML, should occur in the core representation using best practices for rendering those formats.**

**Rationale**

- Utilizing established display conventions from the email environment will enhance users' ability to understand that additional content is embedded, whether or not they are able to explore that additional content.

**4.1.4 The core representation shall include the sender or recipient's email address as well as any alias, when applicable.**

**Rationale**

- Easy access to both address and alias helps users interpret and disambiguate archive contents.

**4.1.5 The core representation should include indicators (e.g., an attachment icon, an exclamation mark to indicate importance, emphasis markup to show whether an email has been read, IMAP flags, etc.) providing information and/or interaction typically available to email client users .**

**Rationale**

- Information normally presented to email users in an email viewer is critical to user acceptance and interaction with EA-PDF archives.
- Maintaining indicators of importance and other such additional data is potentially useful to archive users.

**4.1.6 The core representation shall include the name, format and access to external attachments. Inline attachments should be rendered within the email body. The core representation may include renderings of attachments.**

**Rationale**

- Archive users typically require immediate access to email attachments.
- Users may benefit from renderings (i.e., captured in the core representation) created by the email archive creator

**4.1.7 The core representation should include at a minimum typical header metadata. More detailed and technical metadata shall be available.**

**Rationale**

- Typically, more detailed or technical metadata is not shown by default, but in an archival context, *some* additional metadata may improve usability considerably. For example, displaying the message-id will help with consistent identification and referencing of particular emails.
- Providing a core representation that is similar to typical presentation of header information in an email application improves usability.

**4.1.8 Core representations shall be rendered with standardized page sizes (e.g. A4 or US letter).**

***Rationale***

- Standard page sizes ensure usability and good user experience when printing emails.
- Email standards do not define “pages,” nor are there common conventions for these. This guideline is aimed at reducing the use of arbitrary approaches (for example, having variable page size determined by the length of an individual email) that may hinder standard printing use cases.

## 4.2 Collections of email messages

Email collections may refer to multiple emails from one or more accounts. EA-PDF processors should apply common practices used in email applications to display the organisation and structure of email collections.

There are no specific requirements specifying the order, prominence, or structure of email collections - these can vary from one email application to another and so can be determined by implementers of the EA-PDF standard. However, implementers may wish to consider the following common practices for displaying email collections.

**4.2.1 The structure of email collections into folders or directories should be presented; ideally with the ability to navigate from a folder or directory to a specific email.**

**Rationale**

- The intellectual organization of emails into folders should be maintained when possible to accommodate the creator's original order. Presentation of original order or structure is a common archival principle that improves users' ability to contextualize, interpret, and understand both the collection and individual items within it.

**4.2.2 The email collections should be searchable; with options to search by common fields (such as sender) or free text searches.**

**Rationale**

- An email account or collection can contain a large number of emails; users often rely on full-text search to find individual emails of interest.

**4.2.3 Emails should be sortable by common fields (e.g., order by date sent, or by sender)**

**Rationale**

- Sorting and filtering also enhances the user's ability to find emails of importance when the account or collection is large.

### 4.3 Describing the archive within the core representation

**4.3.1 A summary of the archive, including any justification for or constraints or limits on its creation, should be included in the core representation.**

**Rationale**

- Having a page in the core representation that explains that the file being viewed is an EA-PDF and what is in it will alert users to the fact that the "file" they are viewing is not a simple collection of pages in a document
- Legacy PDF readers may not provide adequate means of indicating the existence of embedded files, metadata or other features of an EA-PDF

## 5 Functional requirements for EA-PDF readers

An EA-PDF reader shall include the capabilities identified in this clause.

### 5.1 Display and search

**5.1.1 EA-PDF readers shall display the core representation of archived email(s); and provide a means of reviewing, filtering, and searching metadata, body content, and (optionally) attachments from the archive as a whole.**

**Rationale**

- Email archives are often large collections of emails (thousands or tens of thousands of individual emails). Search capabilities that leverage structured metadata and provide filtering or other more advanced search techniques will greatly improve the usability of an EA-PDF.

## 5.2 Email extraction

**5.2.1 EA-PDF readers shall allow users to extract individual emails (including all associated header fields, body parts, and attachments) as standalone IMF files.**

**Rationale**

- A wide variety of applications and systems (from generic document management systems to specialized email processing tools) can parse, render, or interpret emails based on the ubiquitous email standards.

## 5.3 Content extraction

**5.3.1 EA-PDF readers shall allow users to extract email components (headers, body parts, attachments) or actual content (text, images, HTML encoding) encoded within PDF data structures for downstream reuse.**

**Rationale**

- Users may wish to work with the PDF data structures or files directly rather than extracted emails in IMF format.

## 5.4 Metadata extraction

**5.4.1 EA-PDF readers shall allow users to extract PREMIS and other archival metadata for downstream analysis and use.**

**Rationale**

- Many archive systems and repositories can parse and make use of structured metadata using these common standards; making it easy to extract the metadata in its intended format enables or increases interoperability with other systems.

## 5.5 Attachment representation and extraction

**5.5.1 EA-PDF readers shall allow users to extract attachments from the core representation.**

**Rationale**

- Individual attachments are often subject to specific curatorial and archival requirements as stand-alone resources.

### 5.5.2 EA-PDF readers may provide access to attachments from the core representation.

**Rationale**

- Users should be able to immediately view any attachment to an email.

# Appendix A - Problems with existing email formats

Email originated decades ago as a method for transmitting relatively simple text messages between computer terminals. However, despite the interoperability of the many email systems in use today, there remains no agreed-upon method for storing and preserving the information contained in what we generically refer to as “email.”

The capabilities associated with email have expanded far beyond simple text messages; additional protocols were developed to provide interaction between email and other systems. Email applications now integrate support for rich content, nickname databases, distribution lists, arbitrary attachments, encryption, digital signatures, calendar invitations, voicemail messages, and linked data. This complexity and intermingling of content types that are all generically described as “email” can make it difficult to identify exactly what should be preserved and which format is most appropriate. Individual messages, message strings, the folders they are placed in by users, calendars, and account-related metadata held in applications but not defined in an email-related RFC can all be considered important but not all formats are capable of retaining this information.

Numerous formats have been developed to store email messages, but when measured against sustainability frameworks such as the Library of Congress’s, none are currently viewed as a perfect solution for preserving email message and account data through time.

Complicating things further, many email systems store messages and related content in a manner similar to a database rather than in binary formats. Component parts are often stored in database tables and folders and are only pulled back together to display, print or export selected messages.[3] The email application defines these mechanisms and the available formats for storage and export. Depending on the originating system, migrating email to a “preservation format” might require a complex process that includes multiple format migrations each introducing the risk that data, metadata, or contextual information could be lost.

[3] Microsoft and Novell both use database technologies to store email on the server. Japp Wesselius, “Exchange Database Technologies,” *Simple Talk* (blog), August 22, 2008, https://www.red-gate.com/simple-talk/sysadmin/exchange/exchange-database-technologies/; Microsoft Docs, “Manage Mailbox Databases in Exchange Server,” February 7, 2020, https://docs.microsoft.com/en-us/exchange/architecture/mailbox-servers/manage-databases; Novell Corporation, “GroupWise 18 Administration Guide - Information Stored in the Post Office,” accessed March 6, 2020, https://www.novell.com/documentation/groupwise18/gw18_guide_admin/data/adm_poa_understand_post_office_info.html.

## Common message storage formats

### PST - The Microsoft Personal Folders File Format (PST)

PST is a fully documented but proprietary standard from Microsoft used to store email and related information in Microsoft applications such as Outlook and Exchange.[4] It is often used to export folders and messages of email but can also include calendars, notes, and contacts. Issues relating to PST files include size limitations, the inability to run antivirus software unless they are opened in Outlook or Exchange, difficulty in verifying fixity information as the files change each time they are opened, and frequently problems with corruption. PST files support a number of different technical protection mechanisms including (somewhat weak) password protection and encryption through data obfuscation.

### MBOX

MBOX is the generic term for a family of related file formats, loosely standardized in RFC 4155 and sharing the .mbox extension, which store all of the messages of an entire folder (not an entire mailbox) in a single database file, with new messages appended to the end of the file. MBOX can capture and retain the relationship between all of the messages in a folder but is not designed to capture other information types such as calendars, contacts, and notes. It is widely supported but, as noted in the Pronom Registry entry for MBOX: “Due to the variety of MBOX structures, it is not currently possible to produce an authoritative signature for the format.”[5] The Library of Congress assessment identifies four distinct variants and lists various problems including corruption and incompatibility.[6] Email software applications are usually designed for one mbox variant and are unable to open the other variants. MBOX files can be corrupted if multiple processes are modifying them simultaneously and so file locking is required. Unfortunately, multiple incompatible file locking approaches exist. Mechanisms for encryption and other DRM options are not defined within the MBOX file structure but potentially, these technical protection mechanisms could be supported through the applications that produce them Because MBOX stores the contents of an entire folder in one file, the size of the MBOX single file can become exceedingly large. Any corruption in the file may affect the ability of certain clients to access individual messages or even the entire folder.

### EML (Electronic Mail Format)

EML is a de facto file representation that conforms to RFC 5322, which defines the IMF or Internet Message Format (IMF) syntax, as well as other related standards. Its adherence to IMF means that it is widely supported and EML files can be read by most email applications. The Library of Congress Format Sustainability assessment notes that “There is no known specification that defines EML as a file format to

---

[4] Microsoft Docs, “[MS-PST]: Outlook Personal Folders (.Pst) File Format,” September 29, 2019, https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/141923d5-15ab-4ef1-a524-6dce75aae546.

[5] The National The National Archives (UK), “PRONOM: Details for MBOX,” accessed March 6, 2020, https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1519.

[6] “MBOX Email Format,” web page, November 17, 2016, https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml.

store email messages on a file system although it is commonly considered to be an extension of IMF as defined in RFC 5322". As a result there is no single source for information on the structure of EML files. The PRONOM registry includes separate entries for IMF and for MIME email. The email task force report (p.43) notes that keeping track of threads and attachments with EML files can be a significant problem making it best suited for storing individual messages rather than folders of email or entire accounts. Like MBOX, EML does not natively support encryption although the applications that produce them may encrypt where they are stored on a file system.

# Appendix B - Privacy and ethical concerns

## Informative

More so than almost any other digital content genre, email triggers legal and ethical concerns regarding curatorial acquisition, long-term preservation management, and subsequent use. Email is deeply entrenched into all aspects of contemporary personal, professional, organizational, and social life, often of a confidential, sensitive, or restricted nature (CLIR, 2018). This significantly complicates arriving at an appropriate balance of legitimate and laudable archival desires for effective long-term retention and (eventual) access with controlling statutory, regulatory, policy, and disciplinary best-practice obligations. All of these issues have a significant bearing on the functional specifications of EA-PDF, on the necessary feature sets of conforming EA-PDF writers and processors, and on the systems and workflows supporting persistent stewardship of EA-PDF content.

A given email message's status regarding sensitivity and disclosure is not inherent to the message itself. Rather, determination of proper privacy levels is inextricably bound up with consideration of the variable social contexts of message production, dissemination, acquisition, management, and use: what may be permissible in one context could be inappropriate and lead to consequential legal, monetary, or reputational harm in another (Shilton et al., 2017). Contextual norms regarding access to online material are dependent upon role, information type, purpose, and access conditions (Nissenbaum, 2011). With regard to email archiving, relevant considerations of role include understanding who the original writer, targeted recipient(s), and (eventual) archival reader(s) are. Considerations for information type include the degree to which email, including external attachments and internal protocol headers, contains personally identifiable information (PII) or other data that implicates obligatory protocols such as FERPA, GDPR, GLBA, HIPAA, SOX, etc. (Flynn, 2008). Considerations of purpose include the individual, scholarly, organizational, or legal context and imperative or discretionary nature of the intended use. Considerations of access conditions include role-based eligibility requirements; restrictions on time, place, and supervision of the access; and the scope of the controlling terms of use, including potentially necessary damage waivers or user indemnification.

As promoted by the widely adopted ISO 14721 OAIS reference model for archival systems and programs, and consistent with long-standing archival practice, in many legitimate cases an external access representation of a preserved resource may be a substantially derivative form of the internal managed representation. Thus, a general principle of email archiving should be to capture the richest and most complete representation possible at the point of acquisition and then redact as necessary at the point of request and retrieval.

Nevertheless, it may be necessary in certain circumstances to affirmatively avoid capturing or making accessible for public (or even restricted use) particular subsets of email content. In these cases, however, for purposes of maintaining and documenting appropriate archival provenance, it is important that an EA-PDF file or reader provide some tangible indication of the missing material, either in the core representation or associated PREMIS metadata. To the fullest extent possible consistent with controlling legal, policy, or ethical considerations, this indication should detail the general nature of the missing data, the time and place of its suppression, and the justification for its removal. For example, the removal of an email address could be documented as "Email address removed at the time of collection [or time of

presentation] due to PII sensitivity concerns." Ideally, these indications should be represented in both human and machine-readable forms.

CLIR (2018), *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives* (Washington: CLIR) <https://www.clir.org/pubs/reports/pub175/>.

Flynn, N. (2008), *Email Retention and Archiving: Manage Electronic Records, Minimize Workplace Risks and Maximize Compliance* (Columbus: ePolicy Institute) <http://usdatavault.com/library/Email_retention.pdf>.

Nissenbaum, H. (2011), "A contextual approach to privacy online," *Daedalus* 140(4): 32-48 <https://www.jstor.org/stable/23046912>.

Shilton, K., Wickner, A., Oard, D. W., and Lin, J. (2017), "Protecting sensitive email: Archival views on challenges and opportunities," Digital Scholarship and Privacy-Sensitive Collections workshop, *Digital Humanities 2017*, Montreal <https://cs.uwaterloo.ca/~jimmylin/publications/Wickner_etal_2017.pdf>.

# Appendix C - Metadata Options

Metadata is an essential part of an authentic and trustworthy archival collection. This is no different with email archives, regardless of what preservation actions are implemented.

The working group started an initial exploration into various metadata standards/schemas and digital preservation projects that could be applied to EA-PDF Archives. This work should be pursued more in a possible Phase 2 of this project. Other issues include the handling of attachments, rights/restrictions, and workflows.

This is not inclusive of all fields/elements available within a standard or model listed, but rather, this serves as a starting point of what could be possible. Options reviewed include the following:

**EAXS (Email Account XML Schema) -** Email preservation schema that aligns with IMF and was co-developed by the State Archives of North Carolina and the Smithsonian Institution Archives. Tools following schema create preservation XML files of email accounts/collections (can be a single email message or many email messages).

EAXS can provide:
*Account name*
*Message body*
*Subject*
*Sender*
*Recipient*
*Date*
*MessageID*
*Email client*

Some applications used to create an XML file following the schema also include date range of the account, number of messages, number of duplicates, hash for each message, and other data in log files.

**PREMIS (PREservation Metadata: Implementation Strategies)** - A data dictionary and XML schema for preservation metadata for digital objects that repositories can use for asset management.
*Note: PREMIS 4 is expected to be released in the future*

At the email account/topic level, the PREMIS entities of Events and Agents could be leveraged to have metadata including actions and actors that document:

*Virus scan*
*Ingestion*
*Original software (email client)*
*Migration software*
*Validation*
*Account Name*
*Agency/Organization*
*Processing Archivist*

PREMIS should be flexible enough to also include the fields noted above of date range of the account, number of messages, number of duplicates, etc.

**Dublin Core (within METS/other) -** Both State Archives of North Carolina and the Smithsonian Institution Archives explored the use of Dublin Core with email collections in separate projects at the account level.

Elements include:
*Title of collection/account*
*Creator of account*
*Date of transfer*
*Email account address*
*Coverage dates*
*Subjects*
*Description of account or position of the account holder*

This could be embedded within the XMP of the PDF (see below).

**Email message properties modified from the inSPECT (Investigating Significant Properties of Electronic Content Over Time) project -** The goal was to give a framework for determining significant properties of a digital object that should be maintained during preservation activities. This project included a detailed appendix on email message properties.
Values include:
*Display name*
*Creation date*
*Send date*
*Message ID*
*Subject*
*Keywords*

**XMP/Advanced Properties present within the PDF itself** - This is based on a small PDF test conversion of a group of email messages during Phase 1 and not in practice yet.

*Email mess age itself*
Fields:
*ModifyDate*
*CreateDate*
*MetadataDate*
*CreatorTool*
*DocumentID*
*InstanceID*
*dc:format*
*pdf:producer*
*pdfx:MailFrom*
*pdfx:MailTo*
*pdfx:MailSubject*
*pdfx:MailDate*
*pdfx:MailTransportHeader*
*pdfx:MailFolder*

*Note: Dublin Core above from SANC and SIA could be added to the XMP*

Container/collection for a group of emails within a PDF
Fields:
*ModifyDate*
*CreateDate*
*MetadataDate*
*CreatorTool*

*Note: Other metadata options not included here could be more appropriate.*