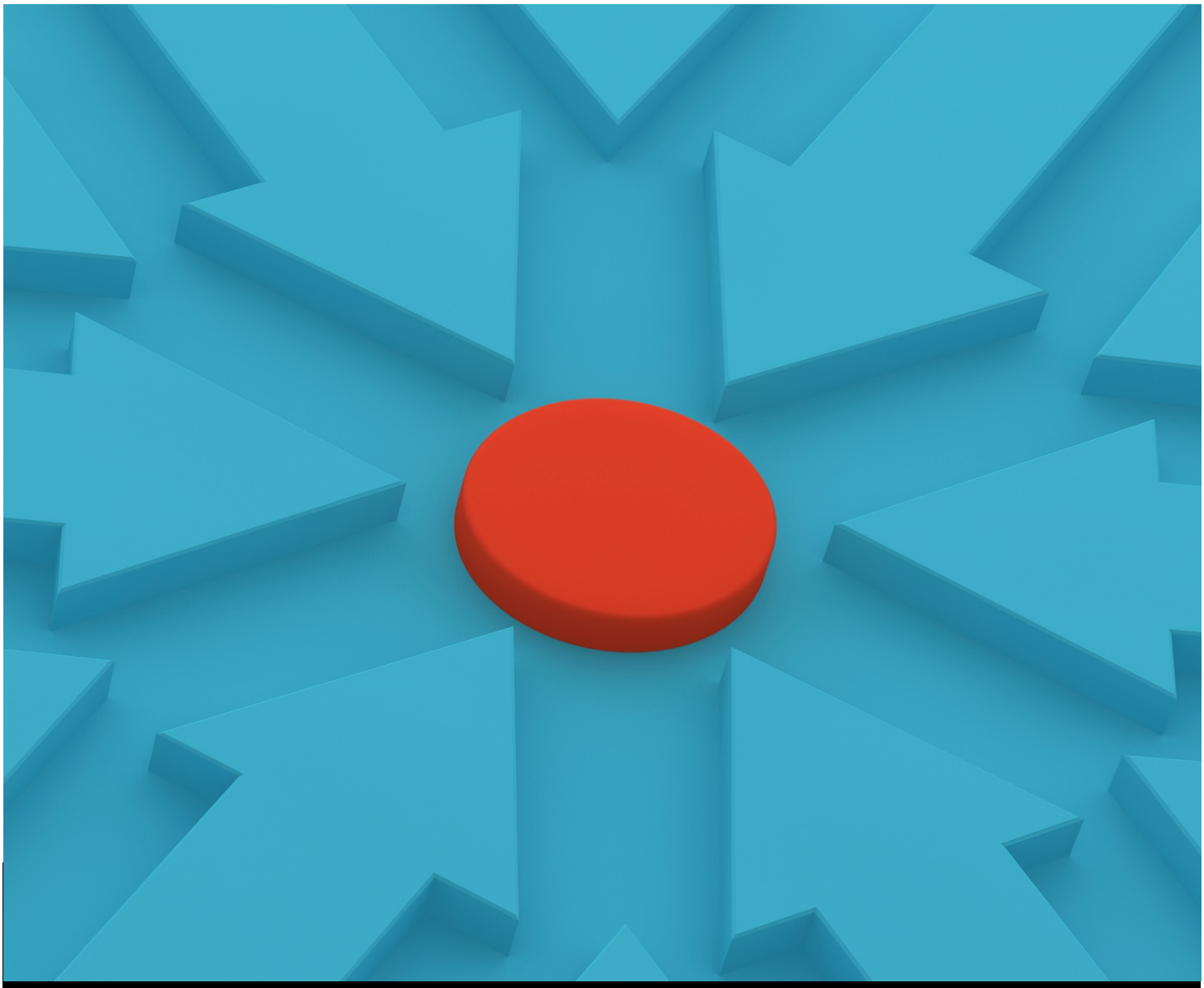
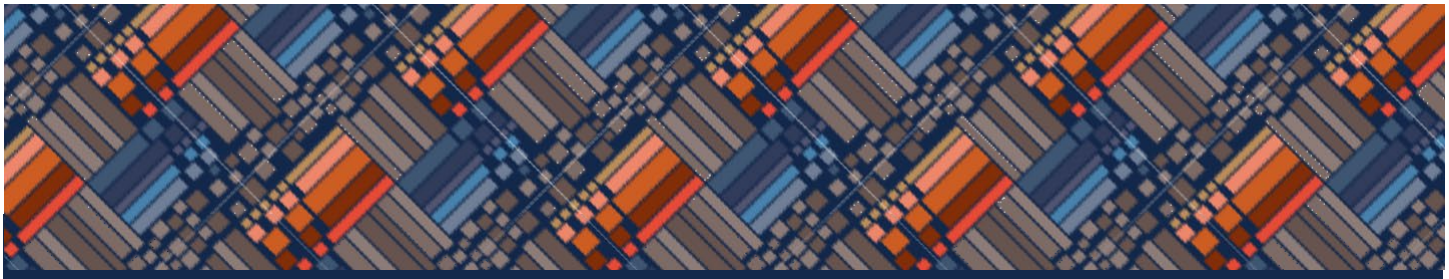


A SPECIFICATION FOR USING PDF TO PACKAGE AND REPRESENT EMAIL

EA-PDF WORKING GROUP

TECHNICAL REPORT PUBLISHED BY
THE UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

TABLE OF CONTENTS



EXECUTIVE SUMMARY

PDF technology is non-proprietary and ISO-standardized. It has a well-known capacity to represent printable documents, and PDF also offers a rich set of data models supporting many types of static content. This document establishes high-level functional requirements for an idealized use of ISO 32000 Portable Document Format (PDF) technology as a model for packaging email for archival or other purposes. These requirements provide a framework within which interested people from the archives, library, museum, digital preservation, and developer communities can collaborate to develop a technically detailed specification and implementation reference model.

Advanced capture, preservation, rendering, and distribution options leveraging PDF technology for email correspondence can deliver long-term value to organizations, communities, and members of the public, while integrating conventional approaches, such as including EML or MBOX content. Simply put, PDF can satisfy rendering and preservation requirements because it is a capable, documented, and open format, supported by an extensive vendor community. PDF is widely implemented; PDF viewers are ubiquitous on consumer and business computers, including handheld devices. In addition, most repositories and digital library applications support PDF. The format allows for the encoding of rich semantic metadata, and can, by way of a usage specification, accommodate interoperable collections of related messages, retaining attachments, email header metadata, folder organization, and server-specific extensions, such as flags and tags. In addition, PDF adds functionality such as redaction, annotation, semantic tagging, and support for digital signatures, all of which are compelling options for various considerations in email archiving.

PDF can thus provide a readily distributable and easily renderable version of an email account, folder or message; a package that is suitable and readily adaptable to downstream educational, business, and personal purposes.

This document describes 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. This document does not specify the manner in which EA-PDF writer or reader software must operate, but does establish a boundary of implementation practice. Further specifications and implementation from this baseline understanding of functional objectives can lead to the development of a proof-of-concept implementation and open source libraries facilitating adoption and incorporation into diverse software applications. An independent, interoperable storage and usage model, operating in complement with other approaches, can cut the Gordian Knot of archiving email.

INTRODUCTION

This document establishes fundamental specifications for archiving email and sets out an approach to considering ISO 32000 Portable Document Format (PDF) technology as a model for packaging email 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

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

- Archivists, librarians, 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, library, museum, digital preservation, and PDF communities can collaborate. We hope that a technically detailed specification and implementation reference model will be developed, using the specifications 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, complementary to other approaches.

Scope

This document establishes requirements and outlines a digital format for capturing and representing email to enable users to exchange and view archived email packages independent of the environment in which they were created or the environment in which they are used.

This document does not specify the following:

- distinct processes for capturing or representing email;
- precise technical design, user interface implementation, or operational details;
- means of storing documents such as media and storage conditions;
- methods for validating the conformance of files or processors;
- requirements for computer hardware and/or operating systems.

This document anticipates that the functional and technical principles described herein will be elaborated into a complete technical specification in a forthcoming (Phase 2) project. ISO standardization is not contemplated or required for EA-PDF until significant industry and user experience has been acquired and integrated into the specification.

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, Comment Period, and Incorporation of Feedback

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

- Commenting on this document using Google Docs comment features: This was 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 was recommended for longer, more substantive comments.

A draft version of this report was posted online in March 2020, and feedback solicited through April 2020. All comments were reviewed by the working group and, where appropriate, adjustments were made to the text.

The working group would like to thank the following individuals for commenting and providing suggestions:

- Matthew Addis
- Hayden Andrew
- Patrick Artner
- Michelle Combs
- Stephanie Decker, University of Bristol, on behalf of Professor David Kirsch, University of Maryland, Dr. Santhilata Kuppili Venkata, The National Archives (UK) Dr. Adam Nix, De Montfort University (UK) and AHRC-funded project "Historicizing the dot.com bubble and contextualizing email archives"
- Andy Harman
- Talj Tatum Harper
- Jan Hutař
- Skip Kendall
- Carol Kussmann
- Tori Maches
- Eveylyn McClellan
- Courtney Mumma
- Amy Rudersdorf
- Arif Shaon
- Lisa Snider
- Megan Sniffin-Marinoff, on behalf of Harvard University Archives
- William Underwood
- Gregory Wiedeman

We also extend our thanks to Ruby Martinez, who provided editorial support and design/layout for the final version of the specification, and to the Andrew W. Mellon Foundation, for providing project support.

OBJECTIVES AND BACKGROUND

Our intention is to set the stage for software developers to create email capture and representation systems leveraging PDF to represent the core metadata, content, attributes, and context that contribute to establishing the digital object's integrity and authenticity while also providing a standardized facility for capturing provenance metadata.

Why Consider Storing Email Using PDF

Email technology does not include a concept of a "native" email preservation format; preservation outside source systems implies some degree of transformation. PDF, on the other hand, is not only accepted by most existing preservation repository structures, it is also a format that is heavily adopted, if not ubiquitous in business and industry, and on the operating systems of most computers and handheld devices.

Email commonly contains personal or business records, and in addition, private and/or vital information. Adoption of archival solutions for email packaging will endure headwinds if they do not leverage existing technologies that are heavily adopted in daily business and personal life. Organizations need to retain email for internal or legal reasons; aligning preservationist needs with business-class solutions offers the possibility of promoting awareness, understanding, and appreciation of digital preservation as an activity.

Present email-to-PDF pathways lack the ability to create effective email archives. For example, printing an email message to PDF from an email client program produces an incomplete version of the message. Most header information is lost, as if you printed the message on paper. While a complete evaluation of current email-to-PDF conversion techniques is beyond the scope of this document, appendix D notes for example, that several significant properties of email are typically lost when converting to PDF from existing software, such as email clients and PDF generation software. Providing an email packaging model leveraging PDF features that support archival needs, would satisfy archival specifications (by, for example, writing detailed header information to metadata fields in the PDF file) and 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. Packaging and representing email with PDF can provide a straightforward, ubiquitous, and highly secure way to access and view archived messages.

Abbreviated EA-PDF Use Cases

This document defines the core archival specifications for capturing email using PDF technology as a container and for presentation purposes. While PDF presents many opportunities to support dissemination, there is no consensus on what the "core" capture dissemination requirements should be. We do not set out hard specifications but do provide some implementation guidance. We recognize that use cases differ. While Appendix D provides a more detailed description of some proposed use cases, abbreviated versions are included here for reference.

"Capture an email account" A server-side administrative option to capture an end user's email account to a single EA-PDF package.

"Archive my email account" A client-side feature to capture the user's email account to an EA-PDF package.

"Archive this email" A client-side feature allowing an end user to create an EA-PDF package containing an archival copy of a single email. This feature could be integrated with workflow software to allow for automatic placement of the file in an appropriate location.

"Extract EML data for processing" An EA-PDF reader extracts the source data from one (or more) EA-PDF packages and places it in a desired location.

"Extract EML data for processing" An EA-PDF reader extracts the source data from one (or more) EA-PDF packages and places it in a desired location.

"Establish provenance" An EA-PDF creator feature facilitating the inclusion of PREMIS or other metadata in machine-readable and core representation forms as well as digital signatures. A matching EA-PDF reader feature would alert readers and provide access to such provenance metadata as the EA-PDF creator provided.

"Find an email in an EA-PDF package using a legacy PDF viewer" EA-PDF packages will offer modest but reliable functionality through typical PDF viewer features such as text search, bookmarks (outlines), links, document structure, annotations and digital signatures.

"Find an email using EA-PDF reader" Although the appearance and precise functionality of the viewer is up to the implementer, viewer developers may compete in providing solutions on the basis of fully-structured data in the EA-PDF package. They may prefer to use the core representation as the foundation of the user experience, or merely as an available node at the end of a search.

In Appendix D, we noted some cases for PDF in packaging email as part of specific, real-life uses, such as the transmission of a user account to a successor or the deposit of email into an existing repository that natively displays PDF files.

Why Email to PDF?

It's entirely fair to ask: "Why would we consider archiving email using PDF at all?" After all, aren't there already capture and migration pathways for email? And why should PDF, of all formats, be considered as a potential packaging format for archival-quality, preservation-enabled emails? These are good questions, and responses may be grouped under two general headings:

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

- Conceptually, an EA-PDF is no more complex than the underlying source email, but represents that complexity in a formally-defined manner, within the structures of the PDF container. MBOX, EML, and other formats that the preservation community sometimes treats as well-defined formats are actually somewhat amorphous families of formats defined more by client implementations than by authoritative specifications.[1] PDF provides a means to represent these implementations in a normalized packaging model, regardless of the underlying source.
- 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 RTF or HTML email content), arbitrary files (e.g., email attachments), source data (e.g., Internet Message Format/IMF), metadata (e.g., header fields), and data to verify authenticity (e.g., digital signatures), all in a machine-readable form.

[1] As Trevor Owens has noted, "Formats are specifications. They are not essential truths about files." In the case of email, format variability between specific implementations is endemic. It can be mitigated to a useful extent by packaging the content into a more formally defined structure, such as the one described in this document. Trevor Owens, *The Theory and Craft of Digital Preservation* (Baltimore: Johns Hopkins University Press, 2018), 120.

- PDF provides a vehicle to capture provenance metadata as part of the act of archiving a mailbox (or server, client, folder, message, etc). Metadata can be written into a "cover sheet" analog, the visual representation of the message or into PDF metadata, using (typically) a container such as XMP.
- Capturing email to a PDF-based format 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.
- Aligning email archiving with business needs facilitates adoption of EA-PDF packages as the standardized means of transfer to and disseminate from repositories. An EA-PDF file would include all component pieces with easy access for typical business needs using standard desktop tools, along with a machine-readable foundation for any other preferred presentation.
- As a single consolidated object within a repository, the EA-PDF package provides another inherent advantage; it is significantly simpler than managing a host of independent but related objects. This is especially true for less sophisticated repository systems or organizations not prepared to support preservation strategies beyond bit-level.

Capturing and packaging email using PDF leverages existing standards and a broad and diverse vendor community.

- There are many use cases for preserving, searching, and reusing email using commodity software and services, such as standard email clients or commercial PDF generation and viewing software.
- PDF facilitates the integration of widely used tools such as email servers and clients into email preservation workflows. For example, such software might produce PDF files that can be easily ingested and displayed in preservation repositories,
- As the *Future of Email Archives Report* notes, "[E]mail archiving is still an emerging practice."[2] Email packaged as PDF could be ingested, stored, preserved, and disseminated via established, widely implemented repository systems that are already in use in government, academic, public, and corporate archives and libraries, many of which natively display PDF files. PDF could supplement other approaches, such as preservation of native email formats (e.g. MBOX, PST, EML).
- Current PDF viewers will render a baseline representation (a complete email message, folder, or even account) without any additional development being required.
- Since the PDF format is highly extensible and widely implemented, a common, published understanding of best practices for archiving email with PDF would facilitate development of viewers capable of presenting such archives with a rich user experience. Such viewers would leverage PDF's native functionality to support extended browsing functions similar to those that exist within email client applications, such as threading, faceting, searching, and sorting by message attributes.

[2] 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.

- There is a good argument to be made for using EA-PDF as the means to transfer and disseminate email to and from a repository, since it packages all of the component pieces in a manner that provides for easy accessibility to an important subset of those components with standard desktop tools. That is a function that would be unavailable if packaging the components in some other container format.
- There is also an argument to be made that managing a single consolidated object within a repository is simpler than managing a host of independent but related objects. This is especially true for less sophisticated repository systems, or organizations not prepared to support preservation strategies beyond bit-level.

In summary, software developers implement solutions that meet a shared understanding of a community's objectives and specifications. This work seeks to leverage existing technologies to allow individuals and institutions a pathway to package and capture email into the most widely used and implemented format for the distribution of text documents.

Other Approaches

Archiving email by packaging and representing it using 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 a packaging 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, such as that supported by EPADD's discovery and delivery modules. Archives, libraries, and other memory institutions should continue to explore these approaches, but to date, they have not widely implemented them as production services. Many archives are simply storing format specific email as unprocessed holdings (for example as system-exported PST or MBOX files that have not been appraised or processed). For these institutions, packaging email as 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.[3]

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 or constraints: convert email to PDF and to do it in a way that preserves the essential provenance metadata that current PDF viewers will render a complete email message, to attest 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, with over thirty years of adoption in business, government, non-profits, and academia, near ubiquitous viewer integration, and a global vendor community supporting its use. In addition, the Library of Congress and

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

National Archives and Records Administration recommend PDF and PDF/A as formats for general archival purposes, for textual documents.[4]

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, first published in 2005. PDF 2.0 (ISO 32000-2) was published in 2017, and will be updated before the end of 2020. ISO standardized and industry-supported PDF technology development is rooted in the not-for-profit and vendor-neutral PDF Association.

PDF Features of Special Interest for EA-PDF

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

- Support for full-text search
- Support for document and object-level metadata applicable to semantic structures
- Support for embedded files, including email source data (e.g., IMF)
- Semantic structures for rich reuse and extraction of content
- Support for redaction, annotation, and linking functionality
- Support for authentication using digital signatures
- Support for packaging provenance metadata
- Support for multi-gigabyte files and millions of pages
- The PDF/A subset specifically targeting archival needs
- The PDF/UA subset specifically targeting accessibility needs
- A high-quality, well-documented, ISO standardized specification supported by tens of thousands of commercial and open source implementations worldwide

Nomenclature

The following terms are normative when used herein; that is, they have specific meanings based on ISO standards terminology. 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

[4] Library of Congress, “Recommended Formats Statement,” web page, accessed July 6, 2020, http://www.loc.gov/preservation/resources/rfs/index.html; National Archives and Records Administration, “Transfer Guidance. Appendix A: Tables of File Formats,” August 15, 2016, https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html.

Terms and Definitions

We use email terms (such as header field, body part, RFC) 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-PDF file.

EA-PDF (package): "Email Archive using PDF" file created according to the provisions of this document

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

EA-PDF reader: interactive viewing software that reads EA-PDF packages

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

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

Embedded file: embedded file stream in a PDF file (see ISO 32000-2: 7.11.4, "Embedded file streams") facilitating inclusion of non-PDF data within PDF files

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

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. And, well-written standards make it easier to develop compliant tools and technology to improve quality and reliability.

Conceptually, EA-PDF is a superset specification leveraging PDF 2.0 (ISO 32000-2), the presumed technical basis for implementing full-featured EA-PDF packages 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), published in November 2020
- ISO 14289-2 (PDF/UA-2), to be published in 2021 or 2022
- ISO 16684-2:2014 (XMP)
- PDF Declarations, The PDF Association, https://www.pdfa.org/resource/pdf-declarations/

1.1. AN EA-PDF PACKAGE 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 PACKAGE 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 PACKAGE 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.

- While accessibility is an important goal, we suggest this is a "should" requirement (and not a "shall" because some accessibility features are difficult or impossible to implement at the point email is captured. For example, 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 PACKAGE SHALL CONSIST OF WELL-STRUCTURED METADATA THAT CONFORMS TO SPECIFIED OPEN STANDARDS.

See Appendix C: Metadata for a non-exhaustive discussion of applicable standards.

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, folder, 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 EA-PDF creators to specify any email messages to be excluded.

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

Rationale

- Although EA-PDF 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, with appraisal or selection decisions being made at a later date.

2.1.2. EA-PDF writers should allow EA-PDF creators the option of retaining 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 EMAIL USING PDF'S EMBEDDED FILES FEATURE

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 organizations) 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 protecting privacy, meeting security or classified information policies, etc. It is beyond the scope of this document to define valid reasons for such exclusions. See requirement 3.1.1 (on recording reasons for exclusions) for further information.
- Header fields are usually short, simple text (thus no strong argument for exclusion due to size or storage) as the data typically requires minimal space relative to the overall body of email data (for example, email attachments often require many orders of magnitude more bytes than header fields).

2.3.2. All captured header fields shall be individually tagged with 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* should be captured

Rationale

- Best practice in archiving email is to capture all body parts and attachments, but this requirement is provided as a "should" requirement because of the potential need to exclude for privacy or security considerations (as discussed in appendix B)
- Retaining all body parts (e.g., plain text and HTML) 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.

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.

2.4.3. Attachments that are themselves emails should be captured as if they were part of the primary collection

- Emails can contain attachments that are themselves emails; it may be desirable to treat these attachments just like every other email in the collection (in other words, applying all of the requirements in this document to those attachments)

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

NOTE: To be clear, the following requirements relate to capturing the content that links *refer* to; we assume that the links themselves are part of the email body (or within header fields) and should be captured like any other aspect of an email body.

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.
- Email can contain links that are intended for tracking purposes (e.g. to track if users have opened an email, read it, etc.), which can be a privacy concern. For this reason it is important to give creators the option to NOT capture any linked content.

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.
- There are many complexities and considerations to capturing linked content (e.g. content that is not retrievable without passwords or within a secured network; content no longer available or changed since the original email was sent, and so on). These considerations are beyond the scope of this report.

2.7. CONSIDERATIONS FOR LEGACY PDF READERS

2.7.1. EA-PDF writers shall be capable of producing EA-PDF packages intended for legacy PDF readers. Fallback functionality shall be made evident to 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 creation options. For example, an EA-PDF reader may include specialized search software processing email metadata in EA-PDF packages, whereas a legacy PDF reader may only include simple text search facilities within conventional PDF files.

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 more accurate and complete understanding of the archive.

3.2. DESCRIBING THE SCOPE OF AN EA-PDF PACKAGE

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 or in accordance with a records retention policy).

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 what was captured at a point in time and provides 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 package was created)**
- **Archive creation software (the software responsible for generating the EA-PDF package)**
- **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 understanding the records themselves.
- This information is, if not generated directly by the software creating the archive, easily obtained and captured by that software.
- 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 packages should include details of the original creation of the email using PREMIS or other similar suitable model/schema in extractable form. If captured, such data shall be captured including, at a minimum:

- **Account holder**
- **Email domain**
- **Institution that hosted that domain**
- **The person, function or entity 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 package.
- 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 packages 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.

This requirement is intended to provide evidence of authenticity from the time that the EA-PDF package is created to when it is taken into custody by an archive or memory institution. If the archival institution is capturing the email themselves or working directly with donors, this may or may not be possible, but is recommended when feasible.

3.4. ADDITIONAL METADATA

3.4.1. EA-PDF writers shall provide a mechanism for users to add additional metadata to EA-PDF packages 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 EA-PDF package, 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 recognizing 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.
- Identifying a standardized use of existing PDF annotation mechanisms to mark up the content of EA-PDF packages 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 user experience of EA-PDF packages will differ between EA-PDF readers and legacy PDF readers. The EA-PDF package's *core representation* (content and layout on a conventional PDF page) provides the basis for a common user experience of the archive and will be an easily accessed dissemination mechanism, given the widespread availability of general purpose legacy PDF readers.

Although EA-PDF allows users to extract messages in a format suitable for email software, the core representation allows users to directly access the archive's content. As a result, institutions preserving email do not necessarily need to maintain additional software for either access or dissemination. Providing the essential elements of emails in the core representation supports backward compatibility with legacy PDF readers, while also allowing for the development of richer experiences with EA-PDF readers.

EA-PDF does not specify the order, prominence, or format of the core representation, as these vary from one email application to another.

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 message body parts of a multi-part message.

Rationale

- The main message content of emails is often provided in multiple formats (most commonly one in plain text 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 plain text is readable, HTML semantics and styling - such as bolded characters or tables - conveys significant information, is more easily readable, and offers a more intuitive usability experience.
- Curators should have the ability to select which formats are displayed by default.

4.1.3. The core representation should display message body parts in rich formats, particularly HTML, using best practices for rendering those formats.

Rationale

- Utilizing established display conventions from the email environment will enhance users' ability to understand content context (e.g., that additional content is embedded quoted or likely unavailable).

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

Rationale

- Email aliases are insufficient for accurate identification of senders or recipients.
- Easy access to both address and aliases helps users interpret and disambiguate archive contents.

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

Rationale

- Information normally presented to email users in an email viewer is critical to user acceptance and interaction with EA-PDF packages.

- Maintaining indicators of importance and other such additional data is potentially useful to archive users.

4.1.6. The core representation shall identify the name and format of any email attachments, and provide end users with access thereto. The core representation should display inline attachments 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 of inline attachments captured in the core representation.

4.1.7. The core representation should include, at a minimum, typical header metadata.

Rationale

- Providing a core representation that is similar to typical presentation of header information in an email application improves usability.
- Immediate access to typical metadata such as From, To, CC and Subject is vital to aid users - especially those with legacy PDF viewers - in consistent identification and referencing of emails.
- While comprehensive metadata is captured within the EA-PDF package (as set out in the requirements in section 3), it's usually unnecessary to display header information in the core representation, if those headers are not also typically used in an email client user interface.

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

Rationale

- Page renderings allow for definitive markup, redaction and other processes.
- Standard page sizes ensure usability and good user experience when printing emails.
- Although 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 good results in use cases that require rendering in printable formats.

4.2. COLLECTIONS OF EMAIL MESSAGES

Email collections may refer to multiple emails from one or more accounts. 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 set as desired by implementers of the EA-PDF standard.

4.2.1. The email collection should be presented as if in its original folders or directories, ideally with the ability to navigate from a folder or directory to a specific email.

Rationale

- The 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 collection 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 that mention a topic or person.
- A user may wish to locate emails with specific attributes, such as to or from a particular person; the ability to limit a search to certain fields enables such searches,

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

Rationale

- Sorting and filtering enhance the user's ability to find emails of importance when the account or collection is large.
- Sorting and filtering may also assist archives staff in creating accurate descriptions of the EA-PDF package.

4.3. DESCRIBING THE MATERIALS INCLUDED WITHIN THE CORE REPRESENTATION

4.3.1. A summary of the EA-PDF package, including any justification for or constraints or limits on its creation, should be included in the core representation.

Rationale

- A page in the core representation describing the purpose of the document alerts users to the fact that this is a special-purpose PDF document, and provides a venue for detailing the archive's intended scope, exclusions, etc.
- Legacy PDF readers may not provide end-users with adequate means of indicating the existence of embedded files, metadata, folder structures, or other EA-PDF features.

5. Functional requirements for EA-PDF readers

An EA-PDF reader shall include the capabilities identified in this clause. These will provide capacities beyond those supported in the core representation or by legacy EA-PDF readers.

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.

Rationale

- Email archives are often large collections of thousands or tens of thousands of individual emails. Search capabilities leveraging structured metadata and providing filtering or other more advanced search techniques will greatly improve the usability of EA-PDF packages.

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 EML 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 IMF standard.

5.3. CONTENT EXTRACTION

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

Rationale

- Users may wish to work directly with extracted emails in IMF format rather than PDF structures. For example, some users may wish to extract certain metadata, text, or attachments to aggregate and analyse in other applications.

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 as well as offering important tools for collection management and analysis.

5.5. ATTACHMENT REPRESENTATION AND EXTRACTION

5.5.1. EA-PDF readers shall allow users to extract attachments in their native formats 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 should provide access to attachments from the core representation.

Rationale

- Users should be able to immediately access any email attachment and view if necessary rendering software is available.

APPENDICES

Appendix A – Problems with Common Email Message Formats

Email originated decades ago as a method for transmitting relatively simple text messages between computer terminals. Despite the relative interoperability of the many email systems in use today, there remains no universally 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 IETF RFC (Request for Comment) 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 that published by the Library of Congress, none are currently viewed as an ideal, or even near-optimal, solution for preserving email message and account data through time.[5]

Complicating things further, many email systems store messages and related content in a wide array of data structures. Component parts are often stored in database tables, with 'folders' only assembled for display, print or to export selected messages.[6] 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.

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.[7] It is often used to export email folders and messages but can also include calendars, notes, and contacts. Issues relating to PST files include size limitations,[8] the inability to run antivirus software unless they are opened in Outlook or

---

[5] Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections. Evaluation Factors, and Relationships," webpage, last updated March 2017, https://www.loc.gov/preservation/digital/formats/intro/format_eval_rel.shtml.

[6] 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.

[7] 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.

[8] Early PST files are limited to 2 GB while files from 2003 onwards can support file sizes up to 20 GB in Outlook 2003 and Outlook 2007 and file sizes up to 50 GB for Outlook 2010 and Outlook 2013. See "Microsoft Outlook PST 97-2002 (ANSI)," web page, November 11, 2013 https://www.loc.gov/preservation/digital/formats/fdd/fdd000377.shtml and "Microsoft Outlook PST 2003 (Unicode)," web page, November 25, 2013, https://www.loc.gov/preservation/digital/formats/fdd/fdd000378.shtml

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[9] and sharing the .mbox extension, which stores 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 unlike PST, it 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."[10] The Library of Congress assessment identifies four distinct variants and lists various problems including corruption and incompatibility.[11] Email software applications are usually designed for one MBOX variant and cannot open the other variants. MBOX files can be corrupted if multiple processes attempt to modify them simultaneously and so file locking is required. Unfortunately, multiple incompatible file locking approaches exist. Mechanisms for encryption and other Digital Rights Management (DRM) options are not defined within the MBOX file structure, nor is a method for redaction; potentially, these technical protection mechanisms could be supported through the applications that produce them, but not necessarily reproduced in the downstream MBOX representation. Because MBOX stores the contents of an entire folder in one file, the size of a single MBOX file can become exceedingly large. As a result, 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 in conformity with RFC 5322, which defines the IMF or Internet Message Format (IMF) syntax, as well as other related standards. EML's adherence to IMF means that it is widely supported; 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 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.

[9] "The application/mbox Media Type", IETF, https://tools.ietf.org/html/rfc4155
[10] The National The National Archives (UK), "PRONOM: Details for MBOX," February 15, 2015, https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1519.
[11] "MBOX Email Format," web page, November 17, 2016, https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml.

Appendix B – Privacy and Ethical Concerns

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 by researchers. Email is deeply entrenched into all aspects of contemporary personal, professional, organizational, and social life, and is often of a confidential, sensitive, or restricted nature.[12] These factors significantly complicate the task of arriving at an appropriate balance between the legitimate and laudable archival desires for effective long-term retention and (eventual) access on the one hand, and applicable statutory, regulatory, policy, and disciplinary best-practice obligations on the other. 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 readers, and on the systems and workflows supporting long-term stewardship of EA-PDF packages and content.

A given email message's status in terms of sensitivity and disclosure is not inherent to the message itself. 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, illegal, or harmful harm in another.[13] Contextual norms regarding access to online material are dependent upon role, information type, purpose, and access conditions.[14] 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.[15] 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 and user or repository 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 accurate 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 of particular subsets of email content, perform redactions at the time of archiving, or facilitate other institutional policies. In these cases, for purposes of maintaining and documenting appropriate archival provenance, it is important that EA-PDF packages and software provide some tangible indication of the EA-PDF creator's choice to exclude certain material, either in the core representation (to guarantee access to legacy PDF viewer users) or associated PREMIS metadata, or

---

[12] Task Force on Technical Approaches to Email Archives, Andrew W. Mellon Foundation, and Digital Preservation Coalition, eds., *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives, August 2018*, CLIR Publication, no. 175 (Washington, DC: Council on Library and Information Resources, 2018),

[13] K. Shilton et al., "Protecting Sensitive Email: Archival Views on Challenges and Opportunities," in *Digital Scholarship and Privacy-Sensitive Collections Workshop* (Digital Humanities 2017, Montreal), https://cs.uwaterloo.ca/~jimmylin/publications/Wickner_etal_2017.pdf.

[14] Helen Nissenbaum, "A Contextual Approach to Privacy Online," *Daedalus* 140, no. 4 (2011): 32–48, https://www.jstor.org/stable/23046912

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

both, as applicable. To the fullest extent possible, consistent with controlling legal, policy, or ethical considerations, and as is typical in US government FOIA workflows, this indication should detail the general nature of the missing data, the time and place of its suppression, and the justification for its removal or redaction. For example, the removal of an email address could be documented as “Email address removed at the time of collection [or “redacted from presentation”] due to PII sensitivity concerns.” Ideally, these indications should be represented in both human and machine-readable forms.

Appendix C – Metadata Options

For every type of digital content, including email, metadata is an essential feature of an authentic and trustworthy archival collection. Email metadata is complex in that access to it depends on who is creating the archive. A message author will likely have access to basic information about the sender name, subject, recipient name, date, file size and whether there was an attachment along with the message body. An account administrator, however, will have access to more detailed information about email clients, transmission and receipt information (such as time sent and time opened) and all addresses including blind copy recipients and nicknames.

The working group that developed this specification identified a few metadata standards/schemas and digital preservation projects that could be applied to EA-PDF packages broadly outlined below, but this work needs to be pursued more in a possible Phase 2 of this project. The working group has not determined which information/fields/elements are essential, nor are these the only options to consider. The key point is that the metadata available to the EA-PDF creator, be they an end user or administrator, should be captured, structured, stored, and represented in a defined manner to enable preservation and use.

EAXS (Email Account XML Schema) - EAXS is a limited framework for preservation that aligns with the IMF format and was co-developed by the State Archives of North Carolina and the Smithsonian Institution Archives specifically for email archiving. It is designed to store XML-encoded message and attachment data for a single email account.[16] Created in 2008 and perhaps in need of being reconceptualized for modern email processing, EAXS was most recently used as part of the TOMES project.[17]

PREMIS (PREservation Metadata: Implementation Strategies) - PREMIS is a data dictionary and XML schema for preservation metadata for digital objects that repositories can use for asset management.[18] While PREMIS does not have specific mechanisms for email archiving, its complex structure of Events (actions which have happened to objects, particularly during their management within a digital repository) and Agents (actors that perform events on the object including the software) could be adapted for this use. One plus for PREMIS is that it is a highly stable, well adopted strategy with robust infrastructure and support across the digital preservation community including an active editorial board. PREMIS 4 is slated for upcoming release.

Dublin Core (within METS/other) - Dublin Core could be an option because of its structured yet flexible nature although, like PREMIS, there is no specific accommodation for email archiving. Current elements could be adapted for this use.[19]

Extensible Metadata Platform (XMP) - XMP is an open, flexible XML-based framework with data models defined by ISO 16684-1:2019.[20] The XMP specification includes many predefined schemas, including Dublin Core, although none specifically for email. Custom schemas can be defined to cover these needs. XMP is already highly supported in many PDF tools.

---

[16] The EAXS schema is maintained in GitHub as part of the TOMES project. https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs, webpage. Accessed July 7, 2020. For a direct link to the schema, use https://raw.githubusercontent.com/StateArchivesOfNorthCarolina/tomes-eaxs/master/versions/1/eaxs_schema_v1.xsd

[17] Transforming Online Mail with Embedded Semantics (TOMES). https://www.ncdcr.gov/resources/records-management/tomes, webpage. Accessed July 7, 2020

[18] PREMIS, https://www.loc.gov/standards/premis/, webpage. Accessed July 7, 2020.

[19] Dublin Core™ Metadata Initiative, https://dublincore.org/, webpage. Accessed July 7, 2020.

[20] ISO 16684-1:2019 Graphic technology — Extensible metadata platform (XMP) — Part 1: Data model, serialization and core properties, https://www.iso.org/standard/75163.html. Accessed July 7, 2020

Appendix D – EA-PDF Use Cases

The following use cases were developed by Ruby L. Martinez, Email Archives Community Fellow at the University of Illinois at Urbana-Champaign.

USE CASE 1: PACKAGE MY EMAIL ACCOUNT

The importance of capturing email has grown exponentially, especially within institutional settings. Prom explains email, in an institutional context, as "an organizational record that documents discussions, decisions, and actions performed in the course of business."[21] Several pathways to archive email have been identified, some of which do not require the use of external software. These allow for email account owners to archive their own personal and business email, either for their own benefit, or to pass on to a successor, or to deposit to an archive.

Let's say I am a retiring museum curator. Over the course of my employment with the institution I have utilized my email account, provided by the institution, as a means to exchange information with donors, related personnel, and others. As a result, I would like to archive my email to preserve this information and pass it along to my successor. To accomplish this, I will outline the steps I would take to archive my email in .PST, .MBOX, and .PDF formats, identifying the benefits and drawbacks of each approach. In addition, I will describe how the process might work with proposed EA-PDF tools.

PST

As a Microsoft Outlook email account holder, I can output my email in .PST format directly from the Outlook application. This approach is the easiest way to archive my email directly from the application and allows me to export email folders and messages, calendars, notes, and contacts. To accomplish this, I will outline the steps below:

1. **Open** Microsoft Outlook
2. Enter **File**
3. Select **Open & Export**
4. Click **Import/Export**
5. Select **Export to a file** and then click **Next**
6. Choose **Outlook Data File (.pst)** and then click **Next**
7. Select the **Mail Folder** that needs backing up and select **Next**
8. Choose a location where to save the backup file
9. Give the file a **name**
10. If you want the file safe from public view, enter and confirm a password and then select **OK**

All the emails within the folder will be saved in the form of a PST file, which can be opened by a wide variety of email applications and converters.

Before looking in detail at the results of this export, it is worth pointing out that Outlook also includes an 'Auto Archive" feature. Its function is slightly different that export, in that it deletes the emails from the server— a feature that may not be apparent to users at first, and that leaves the archive in a potentially vulnerable preservation state, existing only in a PST format on the user's client device.

[21] Christopher J. Prom, "Preserving Email" (Digital Preservation Coalition, December 1, 2011), http://dx.doi.org/10.7207/twr11-01, 11.

Although this embedded feature within Outlook is a simple preservation pathway, the .PST format has some built in limits. For example, larger archives are split into multiple files, and corruption to one part of the file may result in difficulties opening the files. In addition, antivirus programs cannot be easily run against any attachments included in the PST. Similarly, it can be difficult to verify fixity information, as the files may be modified when opened in an email client. However, PST does offer different technical protection mechanisms, including password protection and encryption through data obfuscation.

Once my files are exported, my successor would have to import the files into their own email account or an email client, in order to render and use them. The procedure poses several issues. Importing the files would likely be messy, and it would increase the size of their account and search index, even though the archive is rarely consulted. In addition, there are ethical concerns because of the inability to redact PII or other sensitive information prior to the transfer. Furthermore, there is no way to authenticate the files because I am unable to digitally-sign the resulting archive.

MBOX

In comparison, MBOX, "a generic term for a family of related file formats which stores all of the messages of an entire folder (not an entire mailbox), in a single database file," is another format widely used for capturing email.

However, Microsoft does not currently offer an option to export email as an MBOX file. At this point it would be necessary to reach out to my IT department to convert the PST files I have already downloaded and convert them into MBOX. This can be accomplished through Emailchemy, a commercially-licensed tool used to convert files, or other paid applications that may or may not be supported by local IT support staff.

Additionally, I can take a different approach to converting my own files by downloading BitCurator, software freely distributed under an open source license. BitCurator would provide some processing support by using readpst to convert the PST email objects into MBOX. But again, this may require working with IT and possibly a special exemption to local software policies, to use this open source software.

For these reasons my successor would likely interact with MBOX files similarly to PST files, by importing them into their own email account. MBOX file size can be exceedingly large because MBOX stores the contents of an entire folder in one file. My MBOX files can be corrupted, impeding my ability to access individual messages or an entire folder. Storage, particularly if there are many large attachments and if the server has storage limits, would probably be handled on a client computer, possibly requiring a special backup and preservation strategy. Access would take place through client software, such as Microsoft Outlook, and would be possible only on one device.

PDF (current functionality)

Microsoft also supports a Print to PDF feature. The Print to PDF feature integrated into Windows allows you to save and print any email messages directly as a PDF format. This current email to PDF pathway is time consuming. Although it may be possible to print more than one message to a portfolio, the typical use is to print a single message at a time:

The following steps outline Microsoft's Print to PDF feature:

1. Open your email message that you want to save, and on the File tab, click Print.
2. From the drop-down menu on Printer, choose Microsoft Print to PDF.
3. Choose Print
4. From the Save Print Output box, choose the destination folder; name the file. Then choose Save.

If Outlook doesn't support this feature (it may be because you are not running on Windows 10 or higher), then following steps would be necessary:

1. Open the message you want to save, and from the **File** tab, click Save As option.
2. In the appeared Save As dialog box, from the folder pane, choose your folder where you want to save the file.
3. Select the Save type; as HTML and then click on Save.
4. Now, Open Word and from the File menu choose Open.
5. Select that HTML file you save in Step 3.
6. Save As the selected file as PDF from the file type drop-down list and choose Save.

Once the files are printed, a very limited amount of header information is included in the visual representation of the email, and it is not encoded in the file's properties or metadata. As a result, there is no technical means to audit the file's provenance.[22] In addition, attachment content may or may not be represented in the body of the email and would likely be accessible with secondary software applications. Printing of emails from other programs, such as Apple's Mail app, exhibits similar deficiencies.[23]

EA-PDF (proposed functionality)

With the proposed functionality suggested for EA-PDF creation tools and output, as discussed in this specification, the basic PDF generation process would be similar. Prospectively, a Print to PDF feature integrated to Windows would allow the user to save and print any messages in an EA-PDF compliant format, either as a single message or a group of messages (for example, a folder or an entire account.) Workflow features described above would be the same. Optionally, the print to PDF feature in Outlook would include a set of dialogs, allowing the user to activate or deactivate some of the optional features supported by the EA-PDF format.

An email account converted to the EA-PDF format would have the following attributes or possibilities:

- Preservation of folder structure and thread relationships with an single or multiple related PDF files
- Signature of the package creation attributes (timestamp, software, conversion options, human operator)
- A richer user experience with extended browsing functions outside of email client applications (with use of an EA-PDF viewer)
- A migration pathway independent of external email client applications, promoting the ease of accessibility and dissemination.
- A pathway structured to meet the needs of the preservation community, away from email client implementations, with provenance metadata, email content and attachments, and data to verify authenticity.

[22] Comparing the set of significant properties identified for email in the 2010 InSPECT report (https://figshare.com/articles/InSPECT_Significant_Properties_Testing_Report_Electronic_Mail/7137821) with a PDF representation printed from one or more well-known clients would be a useful exercise, but is beyond the scope of this case study. Such an analysis would show that the PDF is merely capturing the visual presentation of the email in a particular context of a particular client, without provision of significant internal structural and protocol-level metadata.
[23] Duff Johnson and Peter Wyatt, "Hunter Biden's 'Email' and the Potential for Deepfakes with PDF," *PDF Association* (blog), October 19, 2020, https://www.pdfa.org/hunter-bidens-email-and-the-potential-for-deepfakes-with-pdf/.

This use case outlined three packages in different formats, which I could potentially use to hand off to my successor. Current pathways for archiving email typically are stored and preserved closely to their native formats and can typically be viewed with email software, i.e. PST and MBOX. However, this poses an issue when people need to import other's archived email into their own account for various reasons. In addition, current email-to-PDF pathways lack the ability to produce a complete, authentic version of the message since header information is lost. The proposed EA-PDF pathway would significantly mitigate some of these issues, while also requiring the development of new software tools.

For example, proposed functionality of EA-PDF would provide a dissemination pathway independent of email software with core metadata, content, attributes and context intact. Packaging and capturing email into a widely used and implemented format contribute to accessibility because PDF readers are built within most web browsers and operating systems. Users, like my successor, would be presented with a rich user experience and a format they are familiar with.

Use Case 2: Archive a Single Email

Email providers have embedded features in their software that allow for the capturing of a single email message. Each email archiving pathway can differ across the board and not all pathways are capable of retaining account related metadata. Typically, these are print or print-like options, as opposed to export or 'archive' functionality discussed in use case one.

As a graduate student, my initial perception of a "preservation format" for a single email meant exporting my email to my hard drive only for it to be forgotten. Any thoughts of disseminating my email were overlooked. However, I was able to take the necessary step of exporting a single email, unlike many casual users who choose not to do so because of the misunderstanding surrounding the "Archive" functionality.[24]

In short, users are unaware the feature is not a pathway to archive their email but a way to move emails off of the network mail server to free up space in the account. Email archiving pathways can be difficult to implement when there is not a mutual understanding of a "preservation format" or an identified pathway that doesn't require multiple format migrations. Pathways will continue to remain confusing.

The following is a demonstration of exporting an email message from my Outlook and Gmail account and how the files are displayed. In addition, I will address any issues in the two email archiving pathways and how they can be mitigated by the proposed EA-PDF feature specifications.

Outlook

Archiving a single message

1. Open Microsoft Outlook
2. Select email designated for archiving and place in a subfolder.
3. File
4. Open/Export
5. Export to a file (.pst)

[24] From the perspective of a professional archivist, Outlook's "Archive" and "AutoArchive" functionality is at best a misnomer and at worst a hindrance. In essence, the feature removes one or more messages from the server, disconnecting it from the account. Instead, the message or messages is encoded on the local drive (e.g. on a desktop or laptop computer), where it may be accessed through Outlook. Optionally, items may be automatically deleted after an aging date has passed. While this reflects the lay understanding of 'archiving,' it makes the long term preservation and access of the email highly dependent on additional preservation actions. Microsoft Support, "AutoArchive Settings Explained," accessed January 8, 2021, https://support.microsoft.com/en-us/office/autoarchive-settings-explained-444bd6aa-06d0-4d8f-9d84-903163439114.

6. Select Outlook data file
7. Select Archive folder
    a. Select the checkbox to include Subfloders
8. Select option to replace duplicates with items exported
9. Save exported file to a designated location by selecting Browse
10. Click Finish
11. Add password (optional)

After following these steps, a .pst file appeared on my Desktop and can be imported back into Outlook displayed as an Outlook Data File.

Gmail

Archiving a single message

1. Access Google Takeout
2. Under Create A New Export, Select data to include
3. Select option to Deselect all (more convenient if you are only downloading email)
4. Select the check box for Mail
5. Select All Mail data included to select the designated folder to archive
6. Scroll down and select Next Step
7. Select Delivery Method (send download link via email)
8. you can select export once or set up a recurring export schedule
9. Select Create Export
10. Download files via link provided in email

As a result, I have successfully exported emails from both Outlook and Gmail. The step-by-step process to export these varied right down to the exporting file format. Outlook produced a .pst file while Gmail exported a .mbox format file. The respective files can be considered “preservation” format, if I have an email account where I can import them.

Print to EA-PDF

EA-PDF features would simplify the process of packaging email into an acceptable archival copy by circumventing the need for multiple format migrations. Multiple format migrations carry a risk that data, metadata, or contextual information could be lost in the process. The concept of “preservation format” can be lost to casual users of email.

Creating an option for allowing an end user to create an EA-PDF package containing an archival copy of a single email has many implications for casual users and for digital preservation repositories. For instance, migrating email will no longer require a complex process that includes multiple format migrations to be a “preservation format”. The risk that data, metadata, or contextual information could be lost is mitigated by the fact that the email is wholly encapsulated in the PDF wrapper, and that it can be viewed using existing PDF tools.

Embedding an EA-PDF feature into the software would align business-class solutions with preservationist needs and would leverage existing print to PDF functions. PDF has continuously proven

as a reliable solution that is widely used and implemented across businesses, government, nonprofits and academic institutions. In addition, businesses already need to retain email for internal and legal reasons.

Although current pathways allow for retention, there are concerns with maintaining the fixity of email packages. A new EA-PDF pathway run by a system administrator will address these concerns, providing benefits to both businesses and preservation professionals. For example, the proposed workflow will create files that include all component pieces with easy access for typical business needs using standard desktop tools, along with a machine-readable foundation for any other preferred presentation. In addition, since PDF already facilitates the integration of widely used tools and can produce PDF files that can easily be ingested and displayed in preservation repositories, it would provide a dissemination pathway independent of email software. Therefore, the user of EA-PDF packages will have direct access to the archive's content, while also having immediate access to core provenance metadata, content, and attributes, and context within its respective email package. Embedding a print to EA-PDF feature in existing email software holds the potential to garner more clientele because of the impact the feature can have on current workflows. In addition, batch EA-PDF could be integrated with workflow software to allow for automatic placement of the file in an appropriate location.

In conclusion, the proposed functionality of the EA-PDF feature will have great implications and create the opportunity to promote awareness, understanding, and appreciation of digital preservation as an activity.