PDF/MAIL

Moving Theory Towards Practice

Tom Habing
University of Illinois Urbana-Champaign
USA
thabing@illinois.edu

Peter Wyatt
PDF Association
Australia
peter.wyatt@pdfa.org

Eden Irwin
University of Illinois Urbana-Champaign
USA
edeni2@illinois.edu

Ruby Martinez
University of Illinois Urbana-Champaign
USA
rubylm2@illinois.edu

Duff Johnson
PDF Association
Boston,
USA
duff.johnson@pdfa.org

Christopher Prom
University of Illinois Urbana-Champaign
USA
prom@illinois.edu

Abstract – Email is one of the most ubiquitous forms of communication in both personal and professional contexts. The EA-PDF (Email Archiving with PDF) project is developing a PDF specification (PDF/mail) for email archiving, as well as an open-source tool to convert emails to the new PDF format. By creating a new specification defining common understandings for archiving email, the project aims for PDF/mail to lower barriers to effective email preservation that meets the needs of the archive and digital preservation community. This includes building a community to support the project, developing the PDF specification itself, and creating a proof-of-concept tool to convert emails to PDF. With an open-source tool available for converting emails, archivists and other professionals—particularly those working in context where they do not have access to technologies supporting email preservation--will have a straightforward and cost-effective way of preserving emails for posterity, complementing other preservation methods and tools.

Keywords – Email; File Format; Specification Development; Software; Metadata

Conference Topics – From Theory to Practice; We're All in this Together

I. INTRODUCTION

Email is one of the most ubiquitous forms of communication in both personal and professional contexts. Institutions around the world rely on email for all levels of day-to-day operations. However, despite its widespread use and importance as a communication medium, email collections are not being accessioned, processed, or made available in archives at the rate one would expect or hope.

To address this issue, the EA-PDF (Email Archiving with PDF) project is developing a PDF specification (PDF/mail) for email archiving, as well as an open-source tool to convert emails to the new PDF format. PDF technology is already widely used in archives and has many benefits, including ease of use and compatibility with a range of software and hardware.

By creating a new specification defining common understandings for archiving email, the project aims for PDF/mail to lower barriers to effective email preservation that meets the needs of the archive and digital preservation community. With an open-source tool available for converting emails, archivists and other professionals—particularly those working in context where they do not have access to technologies supporting email preservation--will have a straightforward and cost-effective way of preserving emails for posterity, complementing other preservation methods and tools.

This paper provides an in-depth overview of the EA-PDF project and the development of PDF/mail to

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121124

date. This includes building a community to support the project, developing the PDF specification itself, and creating a proof-of-concept tool to convert emails to PDF. This project, and PDF/mail, is working to make email archiving more efficient and accessible and help guarantee emails are preserved before they are gone.

II. Community Development

The project's first main goal is to further goal was to further develop and advance a cohort of individuals from archives and the PDF community, in formal conversation with each other, so that parties with functional expertise in digital preservation, PDF standards development, and PDF technical implementation can iteratively develop the specification. This work centers in the activities of a liaison working group (LWG) hosted by the nonprofit PDF Association. Co-chaired by EA-PDF Project Director Christopher Prom and PDF Association Chief Technology Officer Peter Wyatt, the group has met on a bi-weekly basis since November of 2021. This collaborative structure has been critical to the success of the project, enabling participants with different interests and objectives to work together towards a common goal of developing a specification for using PDF to package and represent email, within the formal specification design process supported by the PDF Association (a non-profit trade group dedicated to providing a vendor-neutral platform for developing open specifications and standards for PDF technology).[1]

The goal of the specification is to provide a clear and comprehensive set of guidelines for the development of the PDF/mail container. The LWG identified several core archival attributes to be included in the PDF/mail container, including defined structures for email data, metadata, and access information, also including a reference copy of the input source file (MBOX or EML). These attributes will ensure that email messages, folders, and accounts would be packaged into archive-ready PDF packages for preservation and reuse.

By utilizing PDF/mail as a standard format for packaging email, this project aims to address the challenges associated with email archiving, such as the lack of simple and accessible email preservation solutions that can be easily adopted by institutions. According to a survey distributed to Illinois repositories, about 60% of respondents indicated that they have not collected any email collections (Martinez et al., 2023). They survey also found that many archives lacked necessary training, technology, and scale of email, were barriers to preserving email. PDF/Mail directly addresses these issues by providing a low barrier solution and building on an already widely used technology.

Accordingly, the project aims to provide a standard and tooling that complements existing approaches but offers a pathway to produce PDF files that fully encode email message metadata, content, and structure, while also allowing for other downstream uses of the files, such as ingest into digital asset management or digital archives software, perhaps where the PDF files serve as an access copy to complement MBOX or EML content that is retained as the preservation copy.

III. File Format Development and Design Considerations

The second project goal is to leverage the community for specification development. Building on the work completed in phase one of the project (EA-PDF Working Group, 2021), the LWG is developing a detailed technical description for the PDF/mail file format, leveraging general-purpose PDF file format features to meet archival needs for preserving and providing access to both the visible content of email messages and message metadata. Files complying with the specification will be usable in today's PDF viewers, (what we term 'legacy' viewers), but will provide a richer navigational experience in software designed for viewing EA-PDF files.

In this way, PDF/mail is similar to ZUGFeRD, Order-X, and Factur-X, all of which use the archival specification for PDF, ISO 19005 (PDF/A) as a foundation and leverage 3rd party standards to define additional domain-specific aspects. However, due to the requirement to preserve source and provenance metadata and email attachments, there is a heavy technical dependence on files embedded inside the PDF/mail file, which requires alignment with (minimally) ISO 19005-3:2012 PDF/A-3 orPDF/A-4f (ISO 19005-4:2020).

[1] https://www.pdfa.org/about-us/

Like PDF/A, PDF/mail includes both file format requirements and a limited set of processor requirements. Due to the variety of possible use-cases, many requirements are expressed as "should" (strong recommendation) rather than hard requirements ("shall"). Like other PDF subset standards, PDF/mail does not define the precise appearance or algorithms that convert an email to PDF, nor does it prescribe content details.

PDF/mail profiles will have three main use cases:

1. A s single email in a single PDF (PDF/mail-1s).
2. Multiple emails in a single PDF (but without a hierarchical or folder-like structure, such as from an MBOX file) (PDF/mail-1m)
3. Container PDFs which contain one or more PDF/mail files, for example, preserving someone's entire email output with various folders, both the usual Sent, Inbox, Draft as well as any other custom organization. (PDF/mail-1c)

At the time of this paper's submission, March 10, 2023, the PDF/mail 0.1 spec was under discussion in the LWG, to be shared more broadly within the PDF, digital preservation, and archives communities in late spring, 2023. The draft specification includes these primary features:

1. PDF/mail files shall be PDF/A-compliant
2. The standard will support metadata describing a corpus of email messages, at the document level of the PDF file, such as name of account holder.
3. At the message level, a set of common email Header Fields are formally categorized as Core Header Fields. The Core Header Fields shall always be present in the "message-level" XMP using Document Part Metadata in in each PDF/mail-1s and PDF/mail-1m file, as well as visually present in the page content of the EA-PDF file using text objects.
4. Where possible, metadata will be mapped to utilize standard Dublin Core metadata fields, such dc:creator for from, or pdf:CreationDate forDate.
5. All PDF/mail-1s and PDF/mail-1m files shall embed the original source email data (e.g., EML, MBOX, OST/PST, etc.).
6. Support for richly formatted email body formats such as HTML and RTF.
7. All email attachments shall be represented inside EA-PDF files as embedded file streams
8. PDF/mail creation software may additionally decide to preserve assets referenced in the source email (e.g., images, SVG), including fetching assets from the internet.
9. PDF/mail permits, but does not mandate, that actionable links in the source email must be link annotations in the output PDF/mail.
10. PDF/mail will allow for preservation of complex hierarchies of folders containing emails, such as Microsoft OST/PST and as represented by many email clients.
11. Support for document structure and navigation features including Tagged PDF and Document Part Metadata (DPM).

In parallel with the specification's development, the University of Illinois is developing an open-source PDF/mail creation tool. The parallel development of the tool has allowed participants in the LWG to react to draft outputs, and for the specification designer to incorporate feedback.

IV. PDF/MAIL TOOL PILOT

Another goal of the project is to produce an open-source, proof-of-concept implementation of an application that could produce PDF files conforming to the new specification. This is available in our GitHub project[2]. This section of the paper provides a high-level overview of that tool.

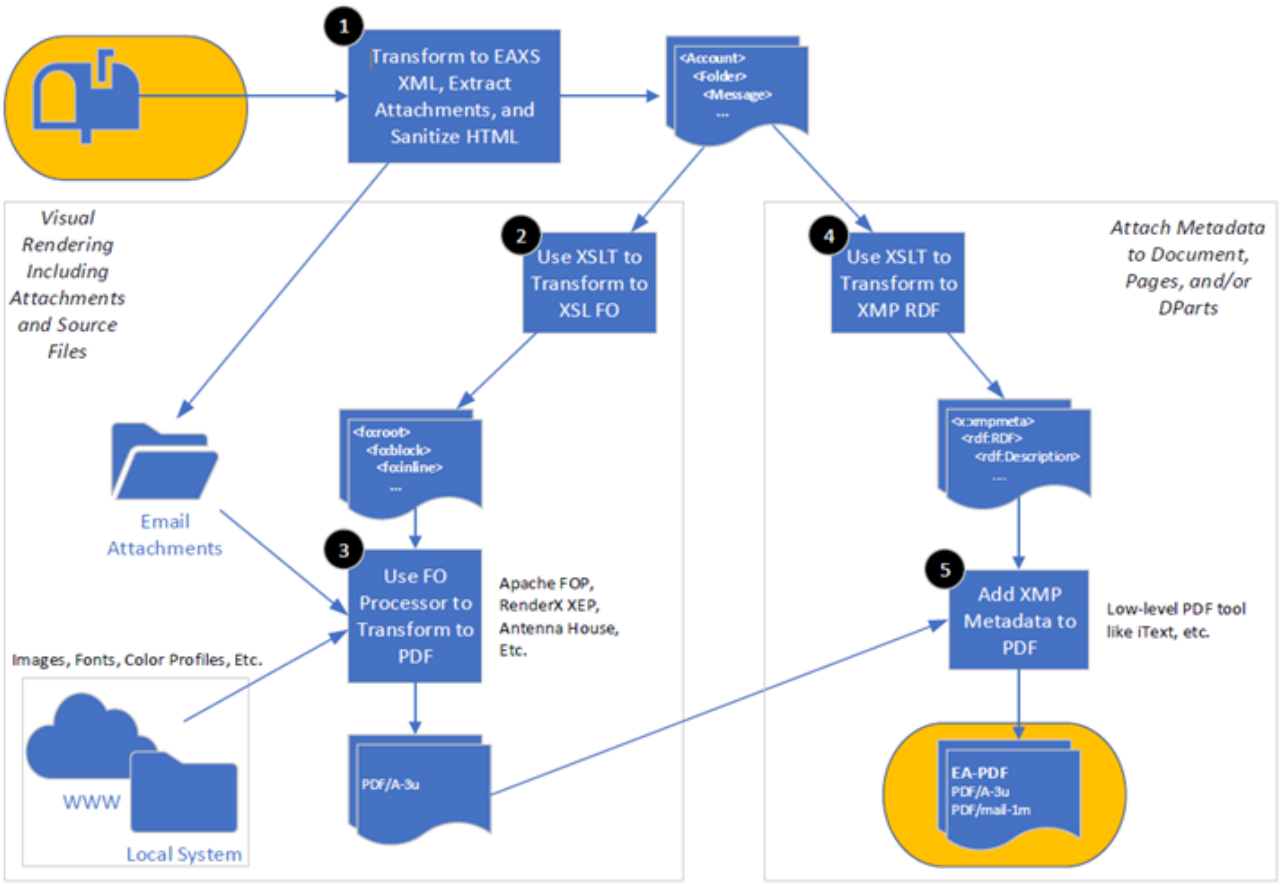
Primarily because of the developer's skillset, the tool was written in the C# programming language using the .NET Core cross-platform framework, allowing the application to be ported to Windows, macOS, or Linux. The tool also utilizes several other open-source libraries, including MimeKit[3] for parsing

[2] https://github.com/UIUCLibrary/ea-pdf

[3] http://www.mimekit.net/

email mbox files, HtmlAgilityPack (HAP)[4] and Fizzler[5] for parsing HTML and CSS, Saxon HE[6] for XSLT transformations, Apache FOP[7] for converting XSL-FO into PDF, ItextSharp[8] for low-level PDF manipulation, among others. The GitHub project includes a basic command-line interface for the conversion tool.

Figure 1. Process Flow for Conversation of MBOX to Archival PDF (EA-PDF)



Referring to Figure 1, Process Flow for Conversion of MBOX to Archival PDF (EA-PDF), there are three major parts of the process: converting the email into XML, visually rendering the XML as PDF, and adding metadata to the PDF. A high-level description of these processes is below; numbers in the diagram above match those used below:

1. Using custom code and the MimeKit, the mbox files are parsed and converted into XML files. The XML schema used for these files is a modified version of the EAXS schema[9] developed for the TOMES project. In addition to creating the XML files, this part of the process also extracts the attachments from the emails; these can be embedded in the XML as base64 encoded data or saved as external files. Any HTML message bodies are also cleaned up and converted to XHTML with inline CSS using HAP and Fizzler; this is done to accommodate the next transformation into XSL Formatting Objects (FO) [10].

2. Using custom XSLT, including a modified XHTML to FO transformation from Antenna House[11], the XML from step 1 is converted into XSL Formatting Objects (FO). The Saxon XSLT engine is used for the transformation.

[4] https://html-agility-pack.net/
[5] https://www.nuget.org/packages/Fizzler.Systems.HtmlAgilityPack: /
[6] https://www.saxonica.com/welcome/welcome.xml
[7] https://xmlgraphics.apache.org/fop/
[8] https://github.com/VahidN/iTextSharp.LGPLv2.Core
[9] https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs/blob/master/docs/documentation.md
[10] https://www.w3.org/Style/XSL/Overview.xml
[11] https://www.antennahouse.com/hubfs/uploads/XSL%20Sample/xhtml2fo.xsl

The XSL-FO is structured with a cover page, and each separate email message is started on a new page, along with a list of attachments at the end of the document. During this step Named Destinations are also added to the FO document for internal linking and so that metadata can be attached to the correct pages as described later in step 5. This step also affords some end-user customization; by modifying the XSLT, the resulting PDF rendering can be altered. It can also be customized to support different open-source or commercial FO rendering engines if desired.

3. Next, using Apache FOP or some other FO processor, the XSL-FO is transformed into PDF. Using processor-specific XSL-FO extensions or configuration settings, the PDF is made PDF/A compliant. The FO rendering engine also pulls the source mbox file and all email attachments into the PDF along with external resources from the web or local file system, such fonts, color profiles, or images linked in the HTML message bodies. This results in a PDF/A document with the significant email message headers rendered as readable text, the plain text and HTML messages bodies rendered as readable text, along with links to the embedded source file and attachments.

4. This step uses another custom XSLT to transform the EAXS from step 1 into XMP RDF metadata. A separate *rdf:Description* is created for each separate email message in the PDF. To the extent possible, predefined XMP properties[12] are used, but some custom properties have also been defined in an extension schema[13] for cases where there is not an equivalent pre-existing property, such as the email headers *to*, *cc*, *bcc*, *in-reply-to*, *references*, etc.

5. In this step the XMP metadata created in step 4 is inserted into the PDF document and linked to the document or to the appropriate page or pages which correspond to the email message described by the metadata. The Document Part (DPart) Metadata (DPM) standard first introduced in the PDF/VT specification[14] is utilized for linking these metadata to the appropriate pages. DPart and DPM allow a set of pages, defined by a start and end page, to be associated with an XMP metadata stream. As mentioned, PDF Named Destinations inserted in the PDF during steps 2 and 3 are used to identify which pages represent which email messages. This step requires low-level manipulation of the internal PDF data structures; the open-source iTextSharp toolkit is currently used for this level of access. In addition to inserting message metadata, this step also inserts document-level metadata, primarily the XML extension schema describing our new non-standard metadata properties. This step can also perform other PDF enhancements that might not be possible using an XSL-FO processer alone (as described in steps 2 and 3), such as adding metadata properties to the attachments, adding watermarks, or setting the default PDF viewer settings like zoom level, etc.

At the end of step 5, the result is an archival PDF/A file which conforms to the new PDF/mail specification. Future enhancements to this tool might include a simple GUI interface for one or more platforms, improved customizations so that end-users can easily change the visual rendering of the PDFs or embellish the metadata with local customizations. Finally, follow-up work should include the development of tools that can render or consume the archival PDF/mail documents for use in

[12] https://www.pdfa.org/resource/technical-note-tn0008-predefined-xmp-properties-in-pdfa-1/
[13] https://www.pdfa.org/resource/technical-note-tn-0009-xmp-extension-schemas-in-pdfa-1/
[14] https://www.pdfa.org/wp-content/until2016_uploads/2011/08/Technical-Introduction-to-PDF-VT.pdf

a digital archive setting, such as user-friendly viewing of metadata, or extracting metadata for searching, categorizing, creating extracts, among many other archival functions.

V. Discussion

PDF/mail is a prospective, under-development solution to a known problem: The need for a simpler, easy-to-use email archiving and access format. Neither the specification nor the tooling described above are intended to offer the only or preferred method to achieve overall repository and institutional needs; digital preservation practice is too complex and varied to support normative solutions. PDF/mail has, in that respect, two goals.

First, PDF/mail is intended to provide the many institutions that have not previously engaged in archiving of email with a low-barrier method to do so. Second, it provides those institutions and many others a distributable, access-forward format that can be accessioned, arranged, described and preserved within existing repository architectures, which often support PDF.

In addition, the PDF/mail proto-standard provides several opportunities for additional research, each of which deserve further exploration, extrapolation, and development.

As noted above, PDF/mail files will include rich, embedded metadata. Looking at this from an archivist's perspective, the PDF Dpart and associated Document Part Metadata reflect archival descriptive practices, which support both hierarchy and other forms of relationships. Document management systems, digital asset management systems, and digital library tools either include the ability to index and harvest embedded metadata or allow developers the ability to implement APIs and other tools to extract and index metadata on input. By providing metadata in a consistent, XMP and RDF-based format, the PDF/Mail standard seeks to enable indexing and discoverability of email messages alongside other digital content.

VI. References

[1] EA-PDF Working Group. "A Specification for Using PDF to Package and Represent Email." Text. Board of Trustees of the University of Illinois, January 2021. https://www.ideals.illinois.edu/handle/2142/109251.

[2] Martinez, R., Prom, C., & Lee, C. (2023, January 13). "Practical" Vs. "Exemplary" Sustainability: Is There a Right Way to Archive Email?_20220914. https://doi.org/10.17605/OSF.IO/6JEBX