

***Email Archiving in PDF (EA-PDF): From Initial Specification to Community of Practice
University of Illinois Urbana Champaign***

Statement of National Need

The Challenge

Why EA-PDF?

Relationship to Other Projects

Project Design

Goals and Outcomes

Focus on Collaboration

Workplan

Success Indicators

Diversity Plan

National Impact

Dissemination

Sustainability

Statement of National Need

[The Future of Email Archives Report](#) noted an outstanding need, nationally and beyond, for archives, librarians and museums to adopt easy-to-implement practices for capturing, preserving, rendering, and distributing email that has continuing value to individuals, organizations, or society (Task Force on Technical Approaches of Email Archives, 2018: 82-83). Accordingly, ***Email Archiving in PDF (EA-PDF): From Initial Specification to Community of Practice*** will (a) foster the development of low-barrier methods to produce authentic usable email packages in PDF format; (b) refine those tools and practices through interactive community feedback; and (c) build long-term, collaborative relationships that are necessary to sustain the conversion of authentic, distributable, and preserveable email packages. By engaging relevant communities of practice in the PDF industry and in government, academic, and other archives—and by working in concert with representatives of other email archiving initiatives—the project will transition a high-level specification to real-world archiving applications.

This 24-month-long project builds upon the recently released recommendation from a planning project that was supported by the Andrew W. Mellon Foundation: [A Specification for Using PDF to Package and Represent Email](#). (EA-PDF Working Group, 2021). That report provides summary recommendations from a multi-institution working group and was refined through community feedback. The project now proposed will fulfill those recommendations by providing tangible outcomes that can standardize preservation-oriented email archiving in the mainstream of archival practice.

Led by the University of Illinois at Urbana-Champaign Library, the project described in this narrative will supply three substantial and concrete products:

- (1) An academic/industry partnership, centered in the activities of an EA-PDF Liaison Working Group (LWG) hosted by the non-profit PDF Association (pdfa.org);
- (2) A detailed technical specification for the proposed EA-PDF (**E**mail **A**rchiving in **P**DF) file format, including requirements for EA-PDF viewing software and implementer guidance; and
- (3) A proof-of-concept, open-source EA-PDF writer.

The Challenge

As a means of communication, email is ubiquitous (Prom, 2018: 4). As a result, it is often the only evidence of a transaction or interaction between individuals. Traditionally, an email message has been accepted as evidence in court if its provenance could be proven to a judge (Pratt, n.d.). Yet email is surprisingly easy to forge. Techniques for altering the metadata used to establish the provenance and authenticity of an email message are common and sophisticated. It is therefore critical that the file formats used to instantiate representations of email outside of their original systems capture and retain the metadata necessary to demonstrate trustworthiness (Johnson and Wyatt, 2020).

Like email, PDF is ubiquitous. Unlike email, it is defined by an ISO standard (ISO 32000) and employed worldwide to capture a wide variety of source document formats in a platform-independent manner. Today, almost every email client includes the ability to save email messages as a PDF file. Unfortunately, none do so in a manner that retains email structures or metadata proving message authenticity. Such outputs are ‘like digital paper:’ versions of the messages lacking an audit trail but that lack many of email’s core attributes (EA-PDF Working Group: 7; Task Force on Technical Approaches for Email Archives: 12). There is a better way.

Why EA-PDF?

Email technology does not include a concept of a “native” email presentation; preservation outside the source systems implies some degree of transformation. PDF, on the other hand, is a format that is broadly adopted for presentation and preservation purposes, prevalent in business and industry, and with viewers installed on the operating systems of most consumer computers and handheld devices.

While emails can be exported, stored, and preserved in something approaching their native formats (for example, PST, MBOX, or EML files), those files are typically rendered and viewed with email software. For security and other reasons, many people will not be comfortable importing others’ archived email into their own email client or system. Likewise, most repository software does not natively display these formats. Well-considered packaging and representation of email using PDF can provide a straightforward, ubiquitous, and highly secure way to access and view archived messages, complementing preservation approaches such as those treated at length in the Future of Email Archiving Report (Task Force on Technical Approaches for Email Archives: 57-75).

While sometimes underappreciated as such, PDF is a natural target format for email preservation. Existing package structures, such as MBOX, reflect application-specific features, and content cannot be easily rendered outside of an email client environment. Domain-specific tools rely on internal databases and are not preservation solutions. PDF, on the other hand, is a supported file format in

most existing preservation repositories and digital libraries. In addition to its familiar page rendering capability, PDF is a highly structured and documented container format supporting dozens of document-specific features and capabilities. PDF technology represents, effectively, a platform-independent free-form database with built-in support for XMP (Extensible Metadata Packaging). These qualities explain its broad appeal and implementation, as well as its suitability for packaging metadata together with visual content. Relevant archives user communities, including local, state, and federal archives, as well as museum archives, university archives, and special collection units, have requested PDF-based archiving options for email (Task Force on Technical Approaches for Email Archives: 82-83).

Invented by Adobe Systems in 1993 and an ISO standard since 2008, PDF technology is broadly adopted worldwide, and benefits from a robust vendor community and mature tools. The format includes support for redaction, attachments, navigation, accessibility, encryption, digital signature technology and more, features that are available to the EA-PDF specification and that can benefit email archiving efforts. Unlike native email formats, PDF includes support for describing the contents of PDF files using XML-based XMP metadata technology. Desktop and mobile viewers are freely available and are typically built into web browsers, making PDF an ideal format for disseminating static information.

Relationship to Other Projects

As noted elsewhere, the community has developed migration-based email archiving workflows, which typically convert email to MBOX or a domain-specific XML format (Smithsonian Institution Archives 2017). These workflows are labor intensive, depending on multiple tools, many of which are specific to the archival community (Task Force on Technical Approaches for Email Archives: 69-74). Conversion to MBOX or XML may not leave an archival repository with a copy of the emails that can be easily rendered or distributed. EA-PDF offers a means to address that gap, and this project seeks to develop a solution that can be integrated with existing email archiving tools, such as ePADD and RATOM (Schneider, Chan, and Edwards 2017; Lee and Woods 2020), with representatives from both projects serving on the advisory board for this one. Likewise, the work to be completed in this project complements the grants being funded by the Email Archives, Building Capacity EA-PDF advisory board and Community Project (<https://emailarchivesgrant.library.illinois.edu/blog/>)

Although libraries and educational institutions are examples of organizations interested in archiving email, government agencies in particular would benefit from a manageable and easily-repeatable email-to-PDF pathway. There are billions of emails at the local, state, and federal level eligible for permanent (or at least long-term) retention residing on email servers and in other systems devoid of substantial digital preservation controls. A comprehensive model for archiving email that was widely available outside of archives-specific workflows would greatly benefit the public. Digital preservation of email records may need to begin before eligibility or transfer to the archives. For example, a 'print to EA-PDF' mechanism could be embedded into email client software or PDF production software such as Adobe Acrobat. In that case, records custodians and managers could produce an authentic, preservation-ready email package with a few clicks of the mouse, or they could automate the process for large volumes of email. Providing a standard package and an easy creation mechanism for those

providing public records through FOIA or other sunshine laws will aid in transparency and documentation of government actions, long before such records reach the archives.

Project Design

Goals and Outcomes

This project will produce three deliverables:

(1) An academic/industry partnership in the form of a PDF Association-hosted EA-PDF Liaison Working Group. As with other industry working groups that develop and support the ISO standards for PDF technology, such as PDF/A and PDF/UA, the PDF Association will provide a virtual work / meeting space for technical collaboration and engagement, including a listserv and administrative support, for development of the EA-PDF specification. The Liaison Working Group will be convened by Chris Prom, the Principal Investigator and Project Director on this grant. Its members will include the Project Co-investigator, PDF Association staff including Duff Johnson and Peter Wyatt, other industry representatives (the Liaison Working Group will be open to all PDF Association members), members of the advisory board for this project, and additional third parties invited by the Principal Investigator and convener.¹

(2) Specification documents documenting the core functionality of the EA-PDF format and the required functionality of creation and rendering applications, with the primary audience being software developers. This includes several elements, first among them a detailed technical file format specification for EA-PDF. Developed within an industry-driven, consensus-based process occurring within the EA-PDF Liaison Working Group, the specification will fully describe file format features necessary to fully capture email messages using the EA-PDF format, aligned with the functional requirements defined in the previous phase of this project. The document will be authored in a manner consistent with other ISO-standardized PDF technologies such as PDF/A.

In addition, the EA-PDF Liaison Working Group will develop a document specifying processor requirements for EA-PDF viewing software along with guidance for implementers.² This document will include discussion of considerations and best practice as well as examples and fragments of PDF

¹ The project's advisory board represents a range of interests and technical or subject matter expertise that will contribute to the overall success of the EA-PDF specification development effort, including individuals from the academic, digital preservation, archival, government community, known to have particular interest in and/or expertise related to email, file formats, or digital preservation. Biographies for advisory board members are included in the supplementary materials to this application.

² While EA-PDF files produced in accordance with the specification will render and be accessible in current PDF Readers (such as Adobe Acrobat Reader or those built into web browsers, the EA-PDF format will define additional functionality that can allowed for an enhanced rendering, browsing, and searching experience, emulating the typical experience of email in client software (EA-PDF Working Group 2021, 22-24).

syntax. As with the technical specification, this document will be authored with software developers as the primary audience, using language and syntax familiar to them.

While ISO standardization of the EA-PDF specification is not immediately contemplated, it is a long-term goal, for potential pursuit after working implementations of the standard have been developed and some degree of adoption and proven marketplace interest has been established.

(3) An open-source proof-of-concept EA-PDF writer hosted by the University of Illinois Library, maintained on GitHub and distributed under the MIT License.³ This open-source tool will extend and refine the DArCMail software, a project previously supported by IMLS, adding EA-PDF to the XML output that DArCMail already supports.⁴ An EA-PDF Application Developer will be hired by the University of Illinois, and this individual will enhance DArCMail so that it produces EA-PDF output consistent with the specification described in section (2) above.

In order to assess and improve tool functionality, the tool will be shared in a public GitHub repository and the user interface will be hosted on a cloud platform (Amazon Web Services). This will provide a scalable means to test the conversion of MBOX files to EA-PDF format. The developer, advised by the PI and Library Information Technology staff, will install and optimize the tool for use in this way, as described in the technology plan that is included as an appendix to this narrative.

A testing rubric will be developed by the Principal Investigator, with input from the advisory board and Liaison Working Group. Through an interactive review process and under the direction of the co-Investigator, the EA-PDF Community Fellow will solicit and test email data from the collections of the University of Illinois Archives and also from the collections held by members of the project advisory board and other archives.⁵ In addition, Illinois will endeavor to make a direct conversion service available, whereby advisory board members and others can self-convert email messages in a secure fashion, with the output returned directly to them, without any intervention by the University of Illinois staff. In this case, the EA-PDF Community Fellow (Graduate Assistant) will schedule meetings with advisors to assess results of the conversion process and report back to the EA-PDF Application Developer and the Liaison Working Group.

³ The MIT License is “a permissive free software license originating at the Massachusetts Institute of Technology (MIT) in the late 1980s. As a permissive license, it puts only very limited restrictions on reuse and has, therefore, high license compatibility.” (MIT License 2021)

⁴ According to the project website DArCMail is Python-based, platform-agnostic and easy to maintain. It has both a command line and a graphical user interface, powerful search and filtering options, which can perhaps be leveraged to add redaction to PDF output (Ferrante, 2018). It has been tested with large volumes of email and based on an assessment conducted by University of Illinois Library staff, could be run on a server to allow integration with other tools and services.

⁵ All email input and output will be transmitted, received, and processed only on secure servers managed by the University of Illinois Library’s Infrastructure and Management Group. Files from non-Illinois institutions will only be made available to their owning institutions and in accordance with University of Illinois security policies. In addition, we will utilize non-disclosure agreements with institutions that request and require them, using either an agreement supplied by the owning institution or the University of Illinois, as appropriate to each circumstance.

Focus on Collaboration

The EA-PDF specifications, tools, and partnerships developed under this grant would complement, not replace, existing investments in other email preservation projects, such as those hosted by Stanford University, Harvard University, the University of North Carolina, and the Smithsonian Institution. For example, implementation of the EA-PDF conversion tools will proceed from a baseline understanding of shared functional objectives represented in the complete EA-PDF specification. This will make it more likely that the proof-of-concept tool may be incorporated into diverse software applications, marking a great leap forward in the ability of archives to preserve and provide access to email, this most intractable of formats.

By engaging relevant communities of practice in government, academic, community, and museum archives, the project will successfully develop a complete technical specification for the EA-PDF file format, build an open source tool to produce valid EA-PDF files, and establish a sustainable partnership with industry members who are interested in developing their own implementations. Both during and following the project's conclusion, the EA-PDF Liaison Working Group will provide the archives and records management community a method to connect with those who develop PDF solutions and software to ensure that EA-PDF creation software is developed on a technically solid, well-documented basis. As such, one possible, even projected outcome of this grant is the availability of both open source and commercial options to produce EA-PDF files— a result that would certainly assist in the massive email management problem that governments, organizations, and individuals currently face.

Beyond the stakeholders represented in the project itself, the benefits of a documented, provenance-preserving, email-to-PDF pathway are not limited to government or academic archives but are applicable across every organization and individual relying on email for daily business. For these reasons, the project includes and is supported by a wide range of collaborators. Confirmed members of the advisory board include representatives the National Archives and Records Administration; the Library of Congress; the Smithsonian Libraries and Archives; Association of Tribal Libraries, Archives, and Museums; Council of State Archivists; Digital Preservation Coalition; North Carolina State Archives; State of Virginia Library; Harvard University; Williams College; Colgate College; Stanford University; University of Albany; and Columbia University.

Workplan

This two-year project will include the following overlapping phases, as specified in the schedule of completion:

1. ***Establish EA-PDF Liaison Working Group (EA-PDF Liaison Working Group)***. Similar to other such technical collaboration and engagement efforts, the PDF Association will host a listserv and workspace including EA-PDF project staff, industry representatives, and archival advisors.
2. ***Create a detailed technical specification for the EA-PDF file format***. Developed within an industry-driven, consensus-based process, the specification will describe file format features necessary to fully capture email messages with relevant attributes using the PDF format

consistent with the functional requirements defined in the previous phase of this project. The specification will be authored in a manner consistent with other ISO-standardized PDF technology standards like PDF 2.0 (ISO 32000-2), with software developers as primary audience.

3. **Develop an open-source proof of concept email-to-PDF writer.** This tool will implement the technical specification and produce files sufficient for its evaluation, testing, and validation. The software will extend the DArCMail open-source software, previously developed by the Smithsonian Institution, to create EA-PDF files. EA-PDF files will be viewable in existing PDF viewers while supporting a full-featured EA-PDF viewing, with an enhanced experience forthcoming in EA-PDF-specific viewers, to be developed after the file format has been fully specified and implemented. (EA-PDF Task Force 2021: 19-21).
4. **Establish requirements and guidance for EA-PDF viewer implementers.** The PDF Association’s EA-PDF Liaison Working Group will define functional requirements for EA-PDF viewers that leverage the full feature set of the EA-PDF format, setting the stage for viewer applications to be developed under a future grant, or by members of the PDF Association who wish to market such tools.
5. **Refine specification and software.** While the Liaison Working Group validates and refines the specification based on implementer review and feedback, project staff at Illinois (Co-Investigator and EA-PDF Community fellow) will test the proof-of concept-tool by converting email collections from the University of Illinois, and other archives— the state, local government, university, museum, and community archives represented among the project advisors.⁶ Testing will verify that the output (a) conforms with the specification and associated PDF format requirements and (b) performs as expected in existing (legacy) PDF viewers. Copies of all converted documents will be supplied to owning institutions, for potential inclusion in processed collections under existing policies and in existing preservation systems. In addition, project staff and advisors will assess prospective integrations with other email archiving applications, such as ePADD or RATOM.
6. **Communicate and Disseminate.** Project PI and staff will partner with the PDF Association and project advisors to present findings and tools at appropriate fora, such as conferences and in publications. Industry members participating in or observing the EA-PDF Liaison Working Group will have the opportunity to assess ways in which the format and viewer specifications may be implemented within their respective commercial software packages. The PDF Association encourages developers to use the PDF Association’s [Solution Agent](#), in order to solicit PDF Association members for OSS licensing of commercial tools. In addition, the Project Director will assess the feasibility for hosting an email-to-PDF conversion service within the “Medusa” repository (Rimkus 2013), which is currently being developed for consortial implementation.

⁶ Collections from the University of Illinois to be converted include administrative records and faculty papers including the email portions of the Charles P. Slichter Papers, Govindjee and Rajni Govindjee Papers, Michael Stern Hart Papers, and Carl Woese Papers, documenting fields such as scientific discovery and open access publishing.

Success Indicators

The following factors are a non-exclusive list of criteria that, when met, would indicate that project goals have been achieved. They will be used to judge the overall success of the project and will be monitored and reported by the Project Director throughout the entirety of the project.

- Completion and publication of specification and requirements documents on timeline provided in schedule of completion.
- Demonstration of MBOX → EA-PDF → MBOX conversion with no loss of fidelity.
- EA-PDF implementation in DArCMail is incorporated into a supported repository infrastructure or processing workflow, at the University of Illinois and other repositories.
- Digital Preservation and software development community members contribute to DArCMail code development via GitHub.
- One or more advisory board members commits to in-house testing of the EA-PDF implementation in DArCMail

In addition, the project includes aspirational, but tangible goals, to see wide implementation of the standard. The following actions can be used to assess this overarching goal:

- PDF Association members vote to publish the EA-PDF specification and associated materials as “industry supported” content available from pdfa.org.
- One or more PDF Association members announces current or planned support for EA-PDF in their open source and/or commercial software offerings.

Diversity Plan

The University of Illinois and this project are committed to an environment that welcomes, cultivates, values, respects and supports the differences between and unique contributions that all people and groups bring to society. This project will include archives, collections, and staff members who reflect a wide range of experiences, backgrounds, and perspectives. There are three core elements to this overall commitment:

(1) Diversity in staffing. The Principal Investigator and Project Director will seek to recruit and mentor an EA-PDF Community Fellow (graduate assistant) who will advance the University of Illinois’ commitments to diversity, equity and inclusion. The position is located in the Office of Digital Strategies, which has an established track record of providing meaningful professional training experience, mentoring, and helping students find appropriate post-graduation placements. For example, the office is currently mentoring two students of color, each of whom have been awarded external fellowships supporting students are advancing the profession’s diversity. We seek to continue that track record with this project. Accordingly, we have structured the project so that the student (under appropriate mentorship) is highly visible with the Liaison Working Group and the archival community. For example, project funds will support the incumbent’s travel to conferences to present project outcomes.

(2) Diversity in Partnerships. We will make a special effort to ensure that the specification development and testing process includes contributions from groups who have historically been subject to discrimination or exclusion from technical projects like this one. Specifically, we will work with the Association of Tribal Libraries, Archives, and Museums to identify and foster reciprocal relationships, so that we can learn alongside indigenous communities that hold email archives. In return, the project, may benefit the tribes (and other community archives) by demonstrating clear preservation pathways that jibe with their existing infrastructures. Our work will build on the principles driving a recently announced grant project relating to Native American Oral Histories held by the University of Illinois (Piwowarczyk 2021). For the EA-PDF project, the President and CEO of ATALM has agreed to help and support tribal connections, deepening the University's engagement in the process of healing relationships that were broken as a result of past actions (Witmer 2020). Advisory Board member Jessika Drmacich (Archivist at Williams College), notes that email collecting is an inclusionary process. In her case, it connects multiple people to an organization and email archiving is part of the Williams College DEI initiatives. An email-to-PDF pathway would regularize collecting from many organizations, committees, student groups, people, and other sources.

(3) Diversity in Access. The outputs of the conversion process will expand collection access to people of many different backgrounds and identities. The average person probably does not have the software available to read and access MBOX files in a secure fashion. They would need to import email archives into their own server or client, an action that would (at best) make them uncomfortable or (at worst) open a gaping security risk. Having a permanent archival solution like EA-PDF creates broader equity of access to essential records. Allowing people to use the technology they already have in hand and that supports accessibility practices increases their ability to audit (e.g. see more metadata) in email records. Features like these can use records and contribute to citizen engagement and confidence in the public record. For example, EA-PDF provided via freedom of information requests would preserve essential system metadata, so that and be more accessible than current formats, allowing diverse populations to authenticate public records and to have confidence in their right to know.

National Impact

In short, IMLS investment in this project will have a disproportionately large and welcome influence in helping government agencies, libraries, archives, museums, and businesses preserve and provide access to email archives. This project's tripartite result will open and hold a space that is needed for transformational change in email archiving. Patterned on the process that supports other PDF subset formats such as PDF/A (ISO 19005-4) and PDF/UA (ISO 14289), the EA-PDF specification will offer a shared understanding for the development of both open-source and commercial software, engaging real-world email archiving applications. This work will complement and extend existing archiving approaches for email, such as the ePADD and RATOM projects. It will provide an industry-supported email preservation pathway and an open-source email-to-PDF writer, producing output that can be integrated into email processing workflows and preserved in the many existing digital repositories that already support PDF.

By providing an entirely vendor-neutral platform for preserving email that meets archival requirements and leverages accepted technology, the project is well-positioned to find rapid adoption throughout the library, academic and preservationist communities. Given the project's early and continuous exposure to industry, it is reasonable to anticipate much broader adoption.

Dissemination

Dissemination is structured into the project via the Liaison Working Group. In addition, we will pursue a defined strategy to make the work of this grant widely known. A project website will be established and maintained by Illinois, sharing news from the project, meeting reports, and announcements. The project director will publicize the work of the group at professional meetings (such as Coalition for Networked Information and DigiPres), and the EA-PDF Community fellow will formally present project results at three conferences. In addition, we will pursue publication of project testing results in journals like *The American Archivist* and in peer reviewed conference proceedings, such as iPRES or JCDL.

As the meeting-place of the PDF industry, the PDF Association is well-positioned to ensure that the PDF technology community worldwide is exposed to the EA-PDF concept and specification. In one sense, the Liaison working group is its own dissemination mechanism since it provides an outlet for the archival and digital preservation community to directly interact with PDF specification developers and potential software implementers. In addition, we will present findings from the work at online PDF association events (this is not budgeted since no conference registration fee would be charged). Finally, project results will be disseminated via the PDF Association website, newsletter, and news releases.

Sustainability

Ultimately, we are setting the stage for this project to bootstrap a community of preservationists and vendors who are dedicated to developing commercial and open-source applications of the EA-PDF format. To achieve this objective, the PDF Association will maintain the EA-PDF Liaison Working Group beyond the funding period, providing a means for project stakeholders and 3rd party implementers and end-users to provide post-funding feedback to the specification writers. This will drive development of updated specifications and improvements to best practice documents. In addition, the University of Illinois will maintain the GitHub repository established to host DArCMail and the code developed under this project. We will explore the possibility of integrating the converter software with other records processes, metadata extraction, indexing, and display tools in our Digital Preservation Repository and Digital Library (Rimkus, 2013), and will encourage others to develop their own integrations.

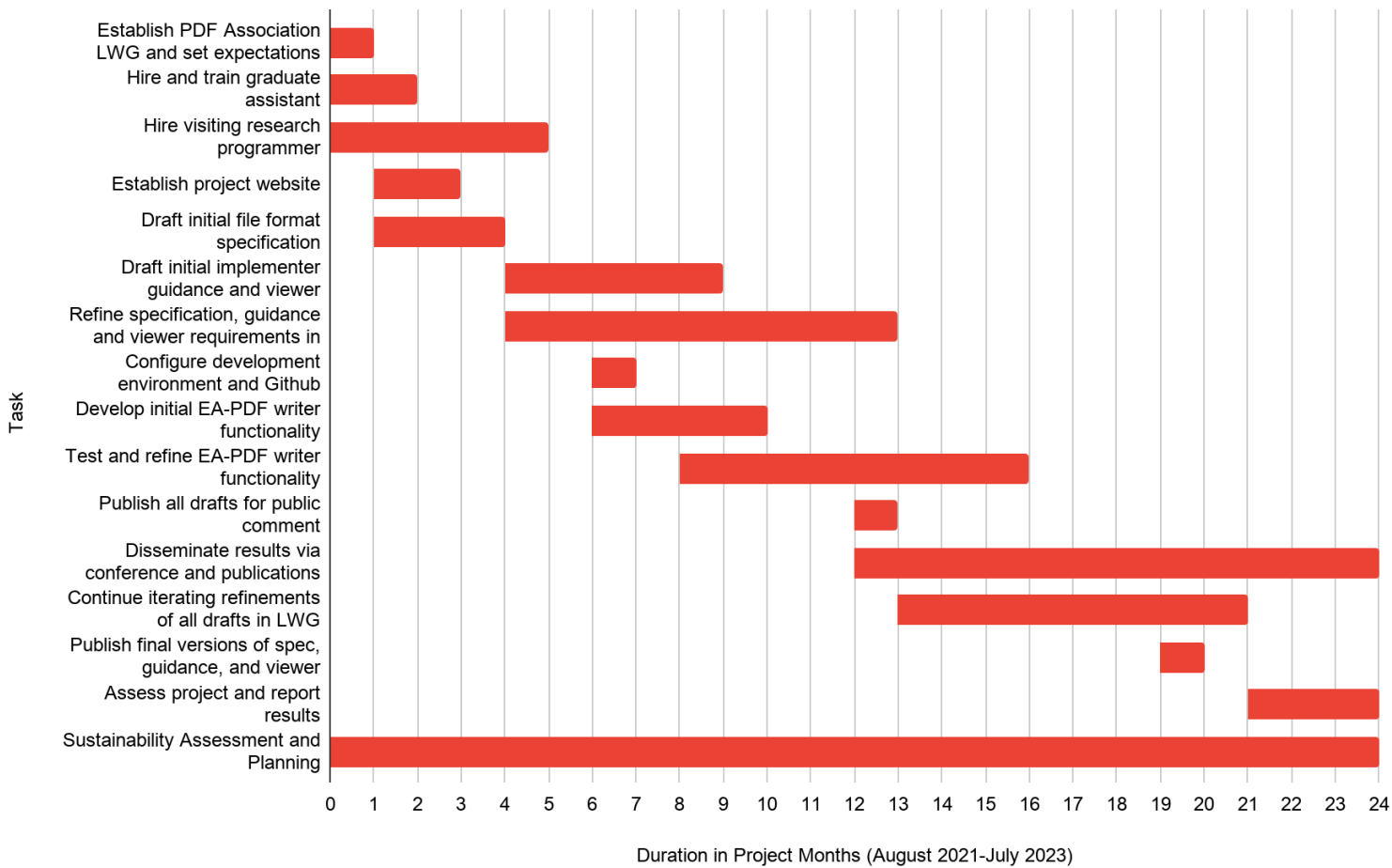
**Email Archiving in PDF: From Initial Specification to Community of Practice
University of Illinois Urbana Champaign**

Schedule Of Completion

As shown in summary form on the chart ([page 2 of this document](#)), the project will proceed on the following timetable. Substantial work will be complete by March 2023, allowing for some flexibility if we encounter unforeseen obstacles or delays. Additional details are provided in the project design section of the narrative.

Task	Start Month	Duration in Months
Establish PDF Association LWG and set expectations	August 2021	1
Hire and train graduate assistant	August 2021	2
Hire visiting research programmer	August 2021	5
Establish project website	September 2021	2
Draft initial file format specification	September 2021	3
Draft initial implementer guidance and viewer requirements	November 2021	5
Refine specification, guidance and viewer requirements in LWG	November 2021	9
Configure development environment and Github repository	January 2022	1
Develop initial EA-PDF writer functionality	January 2022	4
Test and refine EA-PDF writer functionality	March 2022	8
Publish all drafts for public comment	August 2022	1
Disseminate results via conference and publications	August 2022	12
Continue iterating refinements of all drafts in LWG	September 2022	8
Publish final versions of spec, guidance, and viewer requirements	March 2023	1
Assess project and report results	May 2023	3
Sustainability Assessment and Planning	August 2021 (continuous)	24

EA-PDF From Specification to Community of Practice: Schedule of Completion





DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

INSTRUCTIONS

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS**. Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source

code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). Explain and justify any prohibitive terms or conditions of use or access, and detail how you will notify potential users about relevant terms and conditions.

This project seeks to develop technical specifications and software. The specification will be developed under the auspices of a Liaison Working Group within the non-profit PDF Association Inc., whose mission is “to promote Open Standards-based electronic document implementations using PDF technology through education, expertise and shared experience for stakeholders worldwide. Accordingly, the project, PDF-Association, and University of Illinois will make the specifications, and implementation documents developed by this project available with a Creative Commons Attribution License. <https://creativecommons.org/licenses/by/4.0/>. Users will be able to copy and redistribute the specifications and implementation guidance in any medium or format and to remix, transform, and build upon the material for any purpose, even commercially. This will make the outputs of the specification process widely available and facilitate their adoption and translation into open source and commercial software applications.

In addition, the University of Illinois will develop software under this grant. An existing software package, available under the MIT License (<https://opensource.org/licenses/MIT>), will be modified. This is a permissive open source license, ideal for making software from a publicly funded source available for both open source and for commercial developers. Any developer who wishes to reuse or modify the software would be free to do so, even within otherwise proprietary software, provided that they continue to license the underlying DArCmail software and its derivatives with the MIT license.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access

and conditions of use and detail how you will notify potential users about relevant terms or conditions.

This project will convert email collections from the University of Illinois archives and other institutions from a source MBOX format to a derivative EA-PDF format, according to the specifications described in A.1 above.

In the case of the University of Illinois, these materials are already owned under institutional policy or a deed of gift agreement signed by the donor or their representative, maintained in the classified subject file of the University of Illinois and in the files of the University of Illinois foundation. While the project is a proof of concept and does not specifically require or expect that the converted files will be preserved, it is possible that some may be of sufficient value and fidelity to warrant preservation. In these cases, collection specific access policies, complying with the terms of the donor agreement and modeled on other collections, will be developed at the time the files are accessioned into the preservation infrastructure. All terms of access and use will be fully disclosed as part of the collection and finding aid for the parent collection, and access will be equitably managed in accordance with standing policies and practices.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

As with any archival project including email or other records originating in business or personal life, the materials to be processed using open source software may include private information, including but not limited to personally identifying information. In order to protect such records, we will ensure that both input and output files are stored on secure servers (SSL), and in accordance with generally accepted industry standards and University of Illinois policy, where they will be available only to project staff and authorized users from the institutions that own the records.

As noted above, email data will be sent or processed as part of this project, as is common with other archival processing operations or indeed any email system. This data may contain PII or other private information, and will be managed under strict confidentiality processes, complying with the University of Illinois data privacy classification policies and procedures:

https://cybersecurity.uillinois.edu/data_classification

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

As noted in the proposal, this project is a proof of concept that will test the conversion of email records to preservation formats. The files will be stored in MBOX and EML formats, and we anticipate that we will store no more than one terabyte of such data, from no more than 100 different sources or originating email addresses.

A.2 List the equipment, software, and supplies that you will use to create the digital content, resources, or assets, or the name of the service provider that will perform the

work.

As noted in the primary narrative and technology plan, we will use Amazon Web Services (an EC2 instance) to store this data. The University of Illinois has a contract with Amazon Web Services, and the records will be stored under the terms of that contract and University data policies.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

This project will use MBOX (a family of email source formats described by the Library of Congress Sustainability of Digital Format project at <https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml>). In addition, the project will use a new file format EA-PDF, to be developed using the inherent functionality of the existing PDF 2.0 file format, ISO 32000:2 (2020). A summary of the file formats features is available at <https://www.pdfa.org/pdf-2-0-the-worldwide-standard-for-electronic-documents-has-evolved/>

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

As noted in the main proposal, this project is intended to develop a file format specification, build a working proof of concept tool to produce files that comply with the standard, and then compare the output from that tool to the specification, using an instrument to be developed by the Principal Investigator, after the initial draft specification has been prepared. Testing will be an iterative process, involving a combination of manual and automated comparisons, and it will be conducted in close consultation with the PDF Association, acting as contractor who is helping to develop the specification. The EAD Community Fellow will use checklists and testing regimes to compare file output in PDF to the source MBOX. The same file will likely be converted multiple times, as the software is improved. In addition, we will utilize an MBOX -> EA-PDF -> MBOX conversion to assure that all data is being roundtripped with full fidelity.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

This project does not have a specific preservation expectation for the individual files being converted as part of the proof of concept. That said, it is possible that either the University of Illinois or project contributors may wish to preserve the files produced as part of the testing

process. In the case of the University of Illinois, such files would be processed and ingested using our standing procedures, into our preservation repositories, which is compliant with all major preservation requirements (See <https://wiki.illinois.edu/wiki/display/LibraryDigitalPreservation/Medusa+FAQ>)

Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata or linked data. Specify which standards or data models you will use for the metadata structure (e.g., RDF, BIBFRAME, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

As noted in the narrative, all email processed as part of this proof of concept project will be encoded with embedded XMP (Extensible Metadata Platform), instantiating the email header content as EA-PDF file metadata. XMP is an ISO standard, originally created by Adobe Systems Inc., for the creation, processing and interchange of standardized and custom metadata for digital documents and data sets. While this project will not make specific recommendations (other than to suggest that all email headers be preserved as key:value pairs), XMP metadata can be encoded as and/or cross-walked to numerous other metadata formats, such as PREMIS, EAD, or Dublin Core, as suitable for particular implementations. This is because XMP is a generalized container format for metadata, allowing for the declaration of one or more metadata namespaces, not a domain-specific implementation.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

All metadata produced as part of this program will be embedded in the source EA-PDF output, facilitating its long-term preservation within the object of preservation.

If any email collections converted as part of this proof of concept project are assessed for downstream preservation by the host repositories, the use of XMP will provide maximum flexibility in retaining all email header metadata and allowing for its extraction to external search indices and retrieval systems.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

Any metadata to be shared after the conclusion of this project would be subject to the local preservation practices and policies of the owning institution. For the University of Illinois, metadata would be extracted from the EA-PDF files into a custom metadata profile within our digital library, developed specifically for email content and indexing, minimally, the sender, recipient, date, subject, and body of the message. Metadata would be searchable and retrievable in our digital library (subject to access controls and restrictions, which are available down to the item level). Metadata would also be exposed computationally as JSON, allowing for downstream metadata reuse by users or machine agents that are authorized under the access controls.

Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content, delivery enabled by IIIF specifications).

See Answer C3. In addition, any EA-PDF files from the University of Illinois that are produced as part of this project would be rendered using the Universal Viewer, a microservice running as an integral element of our Digital Library (<https://digital.library.illinois.edu/>)

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier for any examples of previous digital content, resources, or assets your organization has created.

While there is no direct analog since this project seeks to develop a new file format, we can point to some sample collections from the University of Illinois Archives, rendered in a browsing interface. We would likely use a similar interface for any outputs from this project that are considered for long term preservation, because many email clients support similar interface controls:

- <https://digital.library.illinois.edu/collections/cd535a80-4852-0133-a7fd-0050569601ca-3>
- <https://digital.library.illinois.edu/collections/53e2b6e0-af03-0132-fd68-0050569601ca-7>

SECTION III: SOFTWARE

General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

This is described in the primary narrative: The proposed software is an open source MBOX to EA-PDF File convertor, based on the DArcMail platform. It will include both a command line and graphical user interface and will take one or more MBOX files (of theoretically unlimited size, subject only to server capacity) as input and produce one or more EA-PDF files as output.

A.2 List other existing software that wholly or partially performs the same or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

As noted, this project seeks to extend the capabilities of an existing software tool, DArcMail (<https://siarchives.si.edu/what-we-do/digital-curation/email-preservation-darcmail>). Since the output to be produced is a new file type, there is no existing software that provides this capability. DArcMail was determined by the Principal Investigator to provide the most promising, lowest barrier platform on which to build such a tool.

Technical Information

B.1 List the programming languages, platforms, frameworks, software, or other applications you will use to create your software and explain why you chose them.

DArcMail is a set of python libraries and scripts. It also uses an embedded SQL Lite database to store and parse data, as part of its processing routines. It exports XML and can also store its output in a SQL database.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

As noted in the main proposal, we are planning this project for a high degree of consistency and applicability to existing email archiving architectures, most of which already ingest or export the MBOX format. Members of these projects are integral to the project advisory board, and will be able to test potential integrations, as the proof of concept proceeds. The microservices and pipeline architecture that DArcMail uses will facilitate rapid testing and deployment in such instances. As a simple implementation, DArcMail could be called by an external application or script to process its output, then pass that output back to the master application or repository.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

N/A

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

As noted in the main proposal, the EA-PDF file format specification will be formally published as part of the existing industry process that led to such ISO standards as PDF/A and PDF/UA.

In addition, the EA-PDF Application Developer will work under the supervision of the Project Director and with advisement from technical staff, to use code documentation tools and best practices for agile code documentation, such as those recommended by <http://agilemodeling.com/essays/agileDocumentationBestPractices.htm>, which are suitable for a proof of concept project and which can be supplemented if the tool is determined to be suitable for mainstream use. In addition, the EA-PDF Community Fellow (graduate assistant) will be mentored by the Project Director in the process of writing effective end-user documentation, and this documentation will also be reviewed by the advisory committee and end users for continuous improvement.

All documentation will be maintained in the GitHub repository established for the project.

B.5 Provide the name(s), URL(s), and/or code repository locations for examples of any previous software your organization has created.

Archon: (retired)

- <https://archon.org>;
- <https://github.com/archonproject/archon>

Preservation Self-Assessment Program:

- <https://psap.library.illinois.edu/>
- <https://github.com/PresConsUIUC/PSAP>
-

Medusa Preservation Repository and associated digital library:

- <https://medusa.library.illinois.edu/>
- <https://digital.library.illinois.edu/>
- <https://wiki.illinois.edu/wiki/display/LibraryDigitalPreservation/Medusa+FAQ>
- <https://github.com/UIUCLibrary> contains many of the underlying libraries that collectively compose the repository and packaging software.

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

It will be provided in a public GitHub repository, under an MIT license. We will welcome and support code contributions from interested parties, including but not limited to the original developers.

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository: TBD

URL: TBD

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

A.1 Identify the type(s) of data you plan to collect or generate, and the purpose or intended use(s) to which you expect them to be put. Describe the method(s) you will use, the

proposed scope and scale, and the approximate dates or intervals at which you will collect or generate data.

N/A

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

N/A

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

N/A

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding, retrieving, displaying, processing, or otherwise reusing the data?

N/A

A.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

N/A

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

N/A

A.7 Identify where you will deposit the data:

Name of repository:

URL:

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

N/A