



NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-260520-OLS-26
Project Category: National Implementation

University of Illinois, University Library

Amount awarded by IMLS:	\$299,845
Amount of cost share:	\$0

The University of Illinois Urbana-Champaign, University Library will seek the full development of open-source tools and libraries for the creation, analysis, and verification of Email Archiving in Portable Document Format files. These tools seek to allow businesses, non-profits, governments, and private individuals to create, view, analyze, and verify files encoded in Email Archiving in Portable Document Format, an archival email format based on Portable Document Format/Archive. Using these software tools, email can be converted to Portable Document Format in a manner that preserves the original messages' fidelity, functionality, and their relationships to other messages in a platform independent format. The objectives of this project are to develop an open-source Email Archiving in Portable Document Format creation and viewing infrastructure; implement software libraries supporting Email Archiving in Portable Document Format file generation and access; establish interoperability with common email storage systems and applications; develop validation tools to ensure Email Archiving in Portable Document Format compliance and authenticity.

Attached are the following components excerpted from the original application.

- Narrative
- Schedule of completion

When preparing an application for the next deadline, be sure to follow the instructions in the most recent Notice of Funding Opportunity for the grant program to which you are applying.

Email Archiving in PDF (EA-PDF): Prototype to Implementation

Introduction

Since before the Declaration of Independence and the founding of the Republic, Americans have preserved correspondence by, to, and about Americans from all walks of life, making these materials available for public inspection, use, and scholarship. Government archives, research libraries, and historical societies are filled with the correspondence files of politicians, thinkers, artists, writers, scientists, and public figures.¹ Preserving correspondence and making it available is an essential democratic function, and the United States was one of the first countries to do so systematically.

But what about now? Who is preserving email, the correspondence for this generation, and how are they doing so?² The Email Archiving in PDF (EA-PDF): Prototype to Implementation Project seeks to ensure that the most ubiquitous correspondence format of the late 20th and early 21st century—email—is as easy to organize and preserve as the paper-based letters and memos of George Washington, Abraham Lincoln, Theodore Roosevelt, Dwight Eisenhower, and Ronald Reagan, or any of their contemporaries and fellow citizens.

Building on work previously funded by IMLS (LG-250129-OLS-21), this twelve-month project will result in the development of open-source tools and software libraries that make it easy to preserve the full content and metadata of email in PDF format. Specifically, these tools will allow businesses, non-profits, governments, and even private individuals to create, view, analyze, and verify files encoded in EA-PDF, an archival email format based on PDF/A. Using these software tools, email can be converted to PDF in a manner that preserves the original messages' fidelity, functionality, and their relationships to other messages, within a platform independent format. This time-limited, highly-efficient project will develop and promote easy-to-use tools and adoption pathways that work for both commercial and open source communities, libraries and archives.

This work is important because it provides a ***tangible and immediate pathway for implementation*** in libraries and archives, and will ***increase public trust*** in institutions that release email content for public inspection. And as a secondary, but extremely important benefit, the EA-PDF will ***jumpstart a commercial market*** for software that is both needed and useful to many businesses, government, academic institutions, and private individuals.

¹ For example, the American Philosophical Society, Founded in 1743 by Benjamin Franklin, preserves some of his papers, alongside others at Yale University and the Library of Congress. Most of these have been published in microfilm, and many are available at <https://www.loc.gov/collections/benjamin-franklin-papers/about-this-collection/>

² In the Society of American Workshop taught by the Principal Investigator of this Project (Christopher Prom), "Email Archiving: Tools, Strategies, and Techniques," the vast majority of archivists attending the workshop worked for institutions that have not preserved any email or are only just beginning their programs.

As detailed in the project results, deliverables include case studies demonstrating the specific functionality needed; software supporting commercial implementations of EA-PDF creation, viewing and validation tools; documentation defining archival best practices; updates to the EA-PDF specification; and news stories, conference presentations, and academic publications.

Project Justification

This project contributes to US business, government, and non-profit effectiveness, while building trust in the same institutions. It also demonstrates the value of both academic research, industry partnerships, and public investments in innovation. Specifically, it:

Demonstrates U.S. innovation at its finest. The Defense Advanced Research Projects Agency (DARPA)³ invented email in the early 1970s along with other networking and telecommunication technologies. It is estimated that over three hundred and seventy-five billion email messages are sent each day allowing for virtually instantaneous communication to all parts of the world (Statista, n.d.). This project builds on a legacy of innovation by investing limited federal resources to solve both an academic, commercial, and government problem: How do we provide access to trustworthy copies of communications that are often sent, stored, and forgotten, but that may have long term historical, cultural or legal value?

Moves email from a dynamic to a fixed state: Today most email is accessed through a web browser. Messages are not stored as individual files. Instead their component parts (To, From, Subject, Date, message body and any attachments) are stored in a database and are only presented together when a user chooses to view or print them. Over time, numerous formats have been developed by vendors to support storing email messages on local computer file systems rather than in database systems, but none are suitable for preserving messages, their component parts, and contextual relationships in a manner that ensures that they remain provably authentic through time (Prom 2019). This project will fix those email messages into a format that demonstrates their authenticity, makes them auditable, and makes them trustworthy.

Solves problems with existing email formats. Existing exchange formats for email are not easily preservable or distributable. The Electronic Mail Format (EML) purports to be an interpretation of the email transmission standard, RFC 822, but no published standard exists (Crocker n.d.). Additionally, EML files can only save individual messages and are not capable of preserving conversations as message threads. The MBOX format is actually a family of related but different interpretations of RFC 4155, which defines the storage database used by MBOX. The various interpretations of MBOX are not fully documented and are not supported by all email applications. Other formats such as Microsoft Outlook Item (MSG), Outlook Personal Folders (PST) by Microsoft Corporation, and Notes Storage Facility (NSF) by IBM Lotus Notes are proprietary and therefore are unsuitable for long-term preservation. In addition, users need to import them into an email client to use them, which is inherently inconvenient and insecure.

³ DARPA is an independent research and development agency within the Department of War (DoW), formerly named the Department of Defense (DoD).

Email applications usually allow for messages to be exported in one of these formats, but each requires the use of a compatible email client for viewing. Viewing an MSG or EML file will usually involve opening it in the user's personal email client rendering it 'live'. Not a desirable result when interacting with evidence in court proceedings or other situations when a user or researcher simply wants to view a message's contents. Additionally, institutional policies often restrict the ability to export messages by end users.

Builds Bridges: This project builds connections within and across multiple communities: government, industry, non-profit and academic. In addition, it will support general improvement in the library field, as it addresses a number of specific issues affecting government, business, and academic archives.

The project seeks to address the following core problems:

PDF files have become a de facto but flawed method to share email publicly. As we saw with the Jeffrey Epstein emails recently released by the Department of Justice, or Hillary Clinton's emails released by the State Department in 2016, PDF is the simplest and most often used method to distribute emails to the public, often via a link on a website. Similarly, PDF is often used when courts or governmental agencies share email records. Converting email to PDF is something that makes sense to most people, and is supported by various 'print to PDF' functions. *However . . .*

PDF's don't convert seamlessly or accurately using current technology. Even if carefully managed, converting messages from a complex communication format to a document format like PDF often results in the loss of significant characteristics and metadata headers. This isn't the result of choosing to save messages as PDF, but simply of the fact that current PDF conversions don't account for email's significant characteristics. Simply put, no one planned for people to convert email to PDF. Attachments, links between related messages, and complete addresses can be lost when converting an email to PDF. As a result, conversions of email messages to PDF are little better than printouts when it comes to attempting to analyze or interact with the data they contain. The standard structural components of email messages are often lost, preventing the data from being easily and accurately searched and analyzed by forensic tools, and undermining the user's trust. And even if one does take a leap of faith to use them, significant amounts of effort are needed to perform any scholarship on this important record type (Task Force on Technical Approaches to Email Archives et al. 2018). *That said . . .*

PDF is actually a perfect container format to fully encapsulate an email message:

Developed by Adobe Inc. in 1993 to ensure that documents maintain a consistent appearance across hardware and software platforms, PDF now supports the inclusion of data types beyond simple text. Today PDF files function a bit like a database, and can include images, audio, video and 3D design files along with extensive structural and descriptive metadata. However, examination of email messages encoded as PDF reveals significant problems since data cannot be written into the format without a structure and software to encode it, which this gran seeks to provide.

Most of the work is already done: The Mellon Foundation funded a pre-study and IMLS funded phase one of this work (EA-PDF Working Group 2021), setting the stage for efficient implementation. Over a three year period, the EA-PDF Liaison working group, hosted by the PDF Association, discussed, developed, and published a technical standard laying out the major and minor features of email messages encoded in PDF (PDF Association, n.d.). In addition, the University of Illinois at Urbana-Champaign developed a proof-of-concept tool to produce EA-PDF files (<https://github.com/UIUCLibrary/ea-pdf>). The application works but is merely a prototype which would not work in a production environment. *Therefore. . .*

A modest amount of additional work remains: The market is primed for practical applications that work in a production environment. Five immediately suggest themselves and are the subject of this application:

- **An open-source library of code**, written in Java, which might easily be integrated into multiple pieces of software. This would drive adoption via closed and open source, setting the stage for PDF companies (such as those belonging to the PDF Association) to innovate and develop their own fully featured applications, while adding additional features that might be attractive to their customers.
- **EA-PDF validation software**, to be used by software developers to test the well-formedness and validity of PDF files created by the creation software.
- **An open source EA-PDF creation tool**. Using the software library noted above, this tool would provide a tangible, easy to use solution for any person or organization to produce basic EA-PDF files and would spur commercial developers to produce more robust solutions. This tool will be hosted by the University of Illinois Library in our paid GitHub repository.
- **EA-PDF viewer code**, which will operate as a plug-in for web browsers such as Firefox and (possibly) Google Chrome and Microsoft Edge. This software is necessary because EA-PDF files leverage capabilities in the PDF/A (archival) format, to show folders, file attachments, and metadata which are not rendered by standard PDF viewers.
- **A free, standalone EA-PDF viewer**, to be used outside of a browser, for longer, more complex files.

This grant fills the needs noted above, because it:

Leverages private industry, government, and university resources. Specifically, the three groups would coordinate and collaborate with the members of the PDF Association's EA-PDF Liaison Working Group to commission five case studies. Managed by the Project Director, the institutions writing case studies will focus on use cases that allow for interactive testing of technology solutions, while communicating user needs and use cases to the developers.

Is based on a known and implemented solution. The development of the EA-PDF specification (based on ISO 19005 PDF/A and released in 2025) solves these problems by defining exactly where in a PDF file the component parts of email messages should be stored, how metadata must be labeled, and how to maintain the association between related messages

(message threads) and attachments. Additionally, the source email message may be included in the EA-PDF file ensuring that the original content and structure are preserved and made accessible. Finally the version of PDF used is PDF/A (archival) document, ensuring that all relevant content is included and that people will be able to tell if it changed since the time it was produced.

Builds on and complements previous private and federal investments. The U.S. government supports innovations related to email preservation. IMLS provided funding to Stanford University Libraries for Email: Process, Appraise, Discover, Deliver – ePADD Phase 2 (LG-70-15-0242-15) and for Phase 1 of this project (LG-250129-OLS-21). Key participants in Phase 1 included representatives from several branches of government including the National Archives and Records Administration (NARA), the Library of Congress (LC), and the Smithsonian Institution (SI). Many people involved in the previous phases will be invited to participate in this project. Its funding will therefore coalesce years of theoretical work into a set of tools that will allow industry, governmental agencies, academic institutions and the public to preserve important records of communication in a hardware and software independent, sustainable format that is already widely adopted world wide.

Opens up scope for digital government: This project will support the move to an “all-digital-government”. Many federal, state, and local governments still maintain workflows and business processes that include paper records. These paper records often begin their lives in digital form but are printed to meet a business need. (Again, we see this illustrated vividly with the Epstein emails, where they were printed to paper, redacted, then scanned to PDF—a very inefficient and wasteful process (Wyatt 2025)). By leveraging the existing PDF standard and ecosystem, the EA-PDF software created by this project will support the use of many of the same tools that are already in place. These include secure digital signatures, password protection and redaction programs, improving the ability of Federal Agencies to respond to Freedom of Information requests in a timely and accurate manner, while reducing the potential for leakage of PII or other sensitive information.

Outputs are “AI ready”: The existing ways to distribute email result in spaghetti-like text dumps of related files, stripped of their context. For example, when MBOX is shared, it is in several, underdocumented formats, specific to particular implementations, such as Google or Microsoft, that may not function reliably in third party tools. Well structured PDF representations of email will be much easier to use with data mining, text mining, topic modeling, generative AI, and other advanced computational techniques that people use to understand complex, messy datasets. The project provides structured metadata in the PDF files that can be input into any AI or other computational environment and reused.

Project Work Plan

Staff and Personnel: This project will benefit from a tight-knit, collaborative core staff and smartly selected consultants and developers, all of whom have an investment in the success of the project.

- **Project Director, Christopher Prom**, is an established leader and convenor of groups working across the library, archives, and digital preservation space. He is a professor and experienced administrator in the second largest academic library in the country. On this project, he will be responsible for setting product vision, convening groups, leading the PDF Association Liaison Working Group, supervising staff, and ensuring timely completion of contracted work. He will also handle all reporting and oversee dissemination efforts. Since he will be on sabbatical during the period of performance of this project, he can pursue this work as a research project contributing to the University's public service mission.
- **Graduate hourly employee and case study manager, Ruby L. Martinez** is a PhD student in Information Sciences at the University of Illinois Urbana Champaign with extensive experience in digital preservation, project coordination, and community-centered archival initiatives. She has contributed to digital preservation training, conference programming, and research dissemination through venues such as the International Conference on Digital Preservation (iPRES) and the Association for Information Science and Technology. On this project, she will assist with coordination, documentation, research synthesis, and community engagement activities.
- **Lead Contractor, ThinkBoxDigital** consists of two former National Archives and Records Administration professionals, Kevin De Vorsey and Greg Lepore, who have over 50 years of collective experience with digital preservation and standards development. De Vorsey participated on the Task Force on Technical Approaches for Email Archives, the EA PDF Taskforce, and is the Convener for ISO TC,171/SC2 WG 5 (PDF/A). Lepore is a globally recognized expert in file format identification and risk assessment. As an electronic records archivist at NARA he oversaw numerous software development projects related to file format analysis, metadata extraction and risk analysis.
- **Five Case Study Developers** will be chosen after application to the Project Director and Lead Contractor, who will jointly organize a peer review process to guide their selection. They will develop EA-PDF specific use cases, to guide the specification updates and development process outlined below.
- **Contractor, PDF Association** is a 501(c)(6) trade association dedicated to maintaining and promoting PDF as a technology, in conjunction with its worldwide members. They are dedicated to a common interest in promoting awareness of PDF's capabilities and best-practice in creating, processing and using portable document format technology. The PDF Association will convene and operate and EA-PDF Liaison Working Group for the purposes of updating the EA-PDF specification to 1.1, will move the updated specification through the PDF Association's publication process and will publishing the new specification on pdfa.org. At a future time, the specification can be considered for adoption through the ISO standards process.
- **Consultant, Big Faceless Organization** is a small corporation which develops Java APIs and tools for working with PDF documents, and is a member of the PDF Association; they develop PDF software for sale to clients that incorporate it into their business workflows. Chief Technology Officer Mike Bremford will consult on this project

and will develop software specifications that will be used by other contractors to develop EA-PDF creation and viewing software.

- **Consultant, TBD.** This will be an established Java programmer with experience in PDF formats, selected via a competitive process, likely from the PDF Association membership. They will develop the open source EA-PDF creation and viewing software applications to the specifications developed by the project.

Phase 1: Project Initiation

Objectives

- Establish project governance and advisory structures
- Re-engage the EA-PDF standards community and stakeholder organizations
- Refine technical specifications and implementation priorities
- Finalize research design and software architecture for EA-PDF tools
- Develop detailed project documentation and dissemination strategy

Key Activities

Project Initiation and Coordination

- Convene the project team and technical staff to confirm scope, deliverables, and timeline
- Establish project communication infrastructure, documentation environments, and version control

Reconvene the Standards and Working Group Community

- Re-engage members of the EA-PDF specification community through coordination with the PDF Association
- Reconvene the EA-PDF Liaison Working Group to review the 2025 EA-PDF specification and identify priorities for implementation support
- Communicate project goals and development plans to relevant communities in digital preservation and records management

Advisory Board Formation

- Establish a multidisciplinary Advisory Board composed of representatives from archives, software development, and government agencies
- Convene the first Advisory Board meeting to review the project plan and provide guidance on implementation priorities and adoption pathways

Use Case and Requirements Development

- Conduct structured discussion with advisory board members and external stakeholders to identify high-priority implementation scenarios
- Document use cases across key sectors, including:
 - Federal and state government agencies responding to records requests
 - Libraries and archives managing email collections
 - Corporate and legal environments requiring evidentiary email records
 - Individual researchers and journalists seeking authentic communication records

Technical Architecture Planning

- Finalize the architecture for the EA-PDF software ecosystem, including: creation tools and libraries, viewing tools, validation and verification tools
- Define interoperability requirements with common email storage formats including EML, MSG, MBOX, and PST
- Establish metadata mapping strategies linking email fields to PDF/A metadata structures

Phase 2: Software Development and Prototype Implementation

Objectives

- Develop the open-source EA-PDF creation and viewing infrastructure
- Implement software libraries supporting EA-PDF file generation and access
- Establish interoperability with common email storage systems and applications
- Develop validation tools to ensure EA-PDF compliance and authenticity

Key Activities

Development of EA-PDF Creation Software

- Design and implement an open-source tool capable of generating EA-PDF files
- Support multiple encoding modes including:
 - Single message per file with embedded original message
 - Single message per file without embedded original message
 - Multiple messages in a single EA-PDF file
 - Collection level EA-PDF files representing entire message sets

Creation of the EA-PDF Conversion Library

- Develop a Java-based open source library to facilitate EA-PDF creation
- Implement conversion routines capable of ingesting email files from multiple formats and platforms
- Ensure preservation of essential structural email components

Development of EA-PDF Viewing Tools

- Develop a standalone viewer capable of opening EA-PDF files without requiring an email client
- Ensure that the viewer displays both document content and structured email metadata
- Create a software library enabling developers to integrate EA-PDF viewing capabilities into other applications
- Ensure compatibility with common PDF viewing frameworks
- Develop an open source validation tool capable of verifying that EA-PDF conform to the specification and archival best practices
- Enable verification of file authenticity and structural integrity
- Enable automated comparison between original email metadata fields and EA-PDF embedded metadata

Specification Feedback

- Provide structured feedback

Phase 3: Integration, Testing, and Implementation

Objectives

- Test EA-PDF tools in real-world archival and records management environments
- Develop best practices and workflows for institutions adopting EA-PDF
- Enable integration with widely used software platforms and web browsers

Key Activities

Prototype Testing

- Conduct extensive testing of the creation, viewing, and validation tools
- Evaluate tool performance using real-world email collections from partner institutions

Institutional Case Studies

- Implement pilot projects with libraries, archives, and governmental agencies
- Document workflows demonstrating how EA-PDF can support: archival preservation, Freedom of Information Act responses, legal evidence management, and scholarly research

Workflow Development

- Development recommended workflows for:
 - Converting email collections to EA-PDF
 - Maintaining authenticity through digital signatures
 - Applying redactions when required
 - Version control and preservation management
- Evaluate authentication methods including digital signatures and provenance frameworks

Browser Integration

- Implement EA-PDF support within existing browser-based PDF viewing frameworks
- Ensure compatibility with major web browsers

Verification Tools for Archival Workflows

- Develop verification routines compatible with archival validation workflows
- Create Preflight verification routines for use within Adobe Acrobat

Phase 4: Documentation, Dissemination, and Adoption

Objectives

- Promote widespread adoption of the EA-PDF format
- Publish implementation guidance and best practices
- Encourage integration into commercial and open-source software ecosystems

Key Activities

Documentation and Best Practices

- Produce detailed documentation covering: metadata mappings, conversion workflows, authentication procedures, and preservation best practices.
- Develop FAQs, tutorials, and technical documentation for both developers and archivists

Implementation Guides

- Publish guidelines for archives, libraries, and government agencies seeking to adopt EA-PDF
- Document recommended practices for: long-term preservation, records disclosure, and secure sharing of email records.

Community Outreach and Training

- Conduct workshops and presentations at professional meetings in the digital preservation and records management communities
- Present EA-PDF tools and findings to organizations

Publication and Research Outputs

- Publish project findings in archival science and digital preservation
- Produce a technical report describing EA-PDF implementation and evaluation

Sustainability and Future Development

- Establish mechanisms for maintaining the open-source EA-PDF tools
- Encourage adoption by commercial vendors and open-source software projects
- Identify opportunities for continued development and community contributions

Project Deliverables

Deliverables include:

- Open source application for the creation of EA-PDF files
- Java library for creating EA-PDF files for integration into existing applications
- EA-PDF support in existing free PDF viewing software
- EA-PDF support for current web browser(s)
- Open source tool and routines capable of verifying EA-PDF files
- Recommendations for updates, enhancements, and bug fixes for the current PDF Association specification for EA-PDF
- Documentation of best practices and optimal workflows
- **Publications and Dissemination:** In addition to the specification and software deliverables, the project members will pursue the following outlets for promoting the results of the IMLS investment:
 - Presentations at conferences such as Society of American Archivists, Association of College and Research Libraries, Coalition for Networked Information, and Digital Preservation Coalition
 - Academic publications such as College and Research Libraries and American Archivist.
 - Exhibits and public lectures as well as national and international news storiss

[illegible]