Email Archiving in PDF: From Initial Specification to Community of Practice

Statement of National Need: The *Future of Email Archives Report* noted an outstanding need for archives, librarians and museums to adopt easy-to-implement practices for capturing and rendering email packages (*Task Force*, 82-83). Accordingly, the Email Archiving in PDF (EA-PDF) project fosters a low-barrier method to produce authentic usable email packages in PDF format, to solicit feedback to further refine such tools, and to build a community of practice for additional work.

PDF is a natural target formation for email preservation. Existing package structures such as MBOX reflect application-specific features and 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 accepted by most existing preservation repository structures.

Fundamentally, PDF is a highly structured and documented container format built on a platform-independent free-form database that includes support for XMP metadata. These features explain its broad appeal and implementation, as well as its suitability for packaging metadata and content. Relevant archives user communities, including local, state and federal government archives, as well as museum archives, university archives, and special collection units, have requested PDF-based archiving options for email.

Project Design: This project builds upon prior work funded by the Andrew W. Mellon Foundation, which resulted in the creation of a "A Specification for Using PDF to Package and Represent Email" (EA-PDF Working Group). This project seeks to implement that specification, which was developed by an expert group and refined after extensive feedback. The project would include the following phases: **Phase 1) August 2021-April 2022:** Development of a minimum viable product, proof-of-concept email-to-PDF writer and associated documentation. Work on the toolset will be completed by the PDF Association under a contract/service level agreement. A system architect will design the detailed technical specification for the creator tool implementation (based on the functional spec noted above), and an independent developer will create associated tooling. The proof of concept will take an MBOX file as input and produce three types of output, depending on user preference: A single email message, a folder or thread of messages, and an entire account/inbox. A second independent implementer will use the documentation to implement round trip conversion, exporting EA-PDF data to MBOX. **Phase 2) May 2022-August 2022:** Testing of this toolset with email collections held by five project partners/archival institutions/sub-awardees—a state archives, a local government archives, a university archives, a museum archives, and a community archives. This phase will include a formal assessment of tool functionality and needs, and will lead to a description of desired features and functionality for both PDF writing/creating and viewing/rendering software, to be used as input for phase three. **Phase 3) September 2022-August 2023:** This phase will lead to the development of specific use cases and a more detailed reference model for next generation EA-PDF creation and viewing, either as standalone applications or integrated into existing software, and dissemination of those in the community. For example, it is anticipated that email specific viewing extensions could improve the searching and browsing experience for email accounts and metadata that have been

packaged in PDF. Work completed during this phase will be led by the University of Illinois, in collaboration with the PDF Association and project partners, shared at conferences and other fora.

Diversity Plan: The University of Illinois and this project are committed to an environment that welcomes, cultivates, values, respects and supports the differences and contributions of diverse people and organizations. This project will include archives, collections, and staff members who reflect a wide diversity of experience, background, and perspectives. We will make a special effort to ensure that these archives include the contributions of groups that have historically been subject to discrimination or a lack of documentation in archives. In addition, the PI will seek to recruit and mentor a student assistant who can advance the university commitments to diversity, equity and inclusion.

National Impact: This project will produce three deliverables:1) a proof-of-concept EA-PDF writer (open source software library); 2) a reference model and documentation detailing the core functionality of EA-PDF creation and rendering applications; and 3) an academic/industry partnership (facilitated by the PDF Association, a non-profit), that will foster the future development and integration of EA-PDF tools into both open-source and commercial software solutions. The EA-PDF (Email Archive in PDF) tools and partnerships developed under this grant would complement, not replace, existing investments in other email preservation projects (Stanford University; University of North Carolina; University of Illinois.) Further specifications and implementation from this baseline understanding of functional objectives, represented in the proof of concept tool can lead to the development of open source libraries and their incorporation into diverse software applications, as well as a great leap forward in the ability of archives to preserve and provide access to this most intractable of formats, An independent, interoperable, and sustainable storage and usage model via PDF, operating in complement with other approaches, can cut the Gordian Knot of archiving email.

Budget Statement: Funding is requested for the following, which would be expended over a two-year period: 1) Support for project director ($1,378 including fringe benefits); 2) Partner subawards ($25,000); 3) Graduate assistant/hourly support to manage partner contacts and writing of specs ($30,845), including fringe benefits and tuition remission); 4) Consultant/Contractor fees to the non-profit PDF Association, to develop a proof-of-concept, open-source EA-PDF writer tool and associated documentation ($130,500); 5) Travel and supply costs, for dissemination ($3,750); 6) Indirect costs ($58,512). Total project budget of $249,985.

Works Cited

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

EA-PDF Working Group. “A Specification for Using PDF to Package and Represent Email.” 28 pages in manuscript. Draft: https://docs.google.com/document/d/1avCVDvMih58NQNMbbH6hR7EzyFeKphOKnMS1QABuE3k/edit?usp=sharing

Stanford University. “EPADD Project Homepage.” Stanford Libraries. Accessed September 24, 2021. https://library.stanford.edu/projects/epadd.

University of Illinois at Urbana-Champaign. “Email Archives: Building Capacity and Community – ".” Accessed September 24, 2020. https://emailarchivesgrant.library.illinois.edu.

University of North Carolina. “Review, Appraisal, and Triage of Mail.” https://ratom.web.unc.edu Accessed September 24, 2020.