Abstract

Stanford University Libraries (SUL), with partners University of Illinois Urbana-Champaign, Harvard University, University of California, Irvine, and Metropolitan New York Library Council, seeks $685,129 in order to develop ePADD (Phase 2), an open-source software package that advances the formation of a National Digital Platform through supporting archival processes around the appraisal, ingest, processing, discovery, and delivery of email archives.

Email offers singular insight into and evidence of self-expression, collaboration, networks, and transactions. Email communications of prominent individuals reveal not only professional and personal actions, decisions, and creative output, but also relationships within society and communities. For these reasons, archival institutions of all types consistently identify email as a critical issue. Yet such institutions face significant impediments to administering email collections, due to concerns about privacy and copyright, and the difficulty of processing large, multi-decade archives with hundreds of thousands of messages. These challenges will only increase, as approximately 246 billion email messages will be sent daily by the end of 2019.

SUL developed ePADD (Phase 1) to confront challenges that donors, archivists, and researchers routinely face in donating, administering, or accessing email collections. ePADD uses natural language processing, automated metadata extraction, and other batch processes to support archival workflows and provide access to otherwise hidden cultural heritage materials. With this grant application for Phase 2 (projected for Nov. 2015 – Oct. 2018), we will direct efforts toward two primary goals: (1) critical functional improvements to the Appraisal, Processing, Discovery, and Delivery modules, and (2) broad and sustained community engagement.

To achieve Phase 2 goals, we will: promote ePADD's integration within an ecosystem of processes and workflows supporting email ingest and preservation; build cross-collection and cross-institution discovery capabilities to improve accessibility of email archives to all users in the United States; facilitate national access to processed email archives approved for public release; advance ePADD's support for restriction and derestriction of materials; optimize ePADD for archives of up to 750,000 messages; and build out new features to augment functionality and performance, incorporating planned additional stakeholder interviews and user testing. We also aim to ensure broad adoption of ePADD through user interface enhancements, as well as extensive community engagement and partnerships. Having partnered with four organizations representing a range of public and private academic institutions, along with cultural and governmental organizations, we intend to create and nurture a user and developer community that is active and enduring. Deliverables will include: the ePADD software program available for free download and optimized to run on modest personal computers; the source code; technical specifications and user documentation; project reports, publications, and presentations; and a project website, serving as a platform for development news and documentation, as well as a community hub.

Satisfying Phase 2 goals will have far-reaching and lasting impact on archival institutions' ability to appraise, process, and provide access to email collections otherwise unavailable to researchers—filling gaps in our national digital capacity. Through Phase 2, ePADD will also be able to build an infrastructure that supports the long-term stewardship and provision of meaningful access to vital cultural heritage materials, as well as foster a community around open-source programs and methods for bulk archival processing—thus equipping repositories with tools to maintain ePADD's adaptability and contribute to its evolution. In this way, fulfilling Phase 2 goals will lay the groundwork for future efforts applying similar automated workflows and functionalities to other classes of born-digital materials.

I. Statement of Need

ePADD is a software package that supports archival processes around the appraisal, ingest, processing, discovery, and delivery of email archives. No other unified tool exists to support all essential components of the archival process, especially the facilitation of donor review, archival processing, and user access. During the past two years (Phase 1), relying in part upon NHPRC funding provided through June 2015, we have proven the concept of using natural language processing (NLP), automated metadata extraction, and other batch processes to support archival workflows and provide access to otherwise hidden cultural heritage materials.

With this grant application for Phase 2 (projected for Nov. 2015 – Oct. 2018), we seek to greatly expand software functionality through critical improvements to the Appraisal, Processing, Discovery, and Delivery modules. This will include promoting integration within an ecosystem of processes and workflows that support ingest and preservation of email; building cross-collection and cross-institution discovery capabilities to improve discovery and accessibility of email archives for all users in the United States; and building out new features to optimize functionality and performance, inclusive of planned additional stakeholder interviews and user testing. We also aim to ensure broad adoption of ePADD through the implementation of UI enhancements as well as extensive community engagement, and partnerships. IMLS NLG funding for Phase 2 would enable ePADD to endure in transforming and improving the discoverability, accessibility, and research utility of email collections. It will also allow ePADD to build an infrastructure to support the long-term stewardship and provision of meaningful access to these important materials.

Archival institutions routinely face significant impediments to administering email collections, including concerns about privacy and copyright, as well as the difficulty of processing large, multi-decade archives with hundreds of thousands of messages (Hangal, 2014; Fung, 2014; for references please see **Supporting Document 1: *Works Cited***). It is not atypical for projects of even modest scope, like NARA's collection of Justice Elena Kagan's email, to require over 6,000 work hours and upwards of 20 archivists and technicians to make available 75,000 individual pages (NARA, 2010). This problem is compounded because email collections can soar to more than a million messages, spanning several decades (HRC, 2013).

The rising demand for an effective solution stems from the research value of email archives: Email offers singular insight into and evidence of a person's self-expression, as well as records of collaboration, networks, and transactions (Zhang, 2015; Sinn, 2011; Pennock, 2006). Email archives further serve as indicia of authenticity and facts beyond the textual communication, as people rely upon email for self-archiving (Sinn, 2011). Email communications of prominent individuals, including politicians, artists, scholars, and the like, reveal not only their professional and personal actions, decisions, and creative output, but also relationships within society and communities (Lukesh, 1999). Thus, the appeal of email collections extends beyond historians to all manner of researchers, journalists, and the general public seeking to obtain insight into individuals and their transactions. This is, in part, why email is the most frequent content type requested under the Freedom of Information Act (FOIA), and increasingly making news (Hawkins, 2014). The potential loss of Hillary Clinton's email, for instance, has created concern that the public could be deprived of a full view of her tenure as Secretary of State (Nicholas, 2015).

Consensus on the need for an effective tool to support archival workflows around email has continued to grow throughout development on ePADD Phase 1. Currently, more than 205 billion email messages are exchanged per day, with a projected annual growth rate of 3%, resulting in an expected 246 billion per day by the end of 2019 (Radicati, 2015). Email is also being sent by a large percentage of the global population, with approximately 2.6 billion email users worldwide (Radicati, 2015). By the end of 2019, it is expected this number will surge to 2.9 billion users—or more than one-third of the population. At the same time that their use and value is increasing, email and digital attachments face mounting risk of deterioration through hardware and software obsolescence, bit-rot, and media failure—threatening our cultural heritage (Kirschenbaum, 2010).

Because use of email is so ubiquitous, and because email records provide such insight into civic, artistic, and scholarly undertakings, archival institutions increasingly seek to acquire email collections (Howard, 2011). The British Library, for example, recently accessioned 40,000 email messages of poet Wendy Cope (Wright, 2011); Stanford University has acquired more than 50,000 email messages of poet Robert Creeley, and over 650,000 of computer scientist Terry Winograd. The Harry Ransom Center has acquired over 17 years' worth of author Ian McEwan's email messages, along with the email archive of McSweeney's publishing, including correspondence with authors like David Foster Wallace and Salman Rushdie (Brown, 2014; HRC, 2013). Indeed, at least 84% of university archives now collect electronic records including email (Noonan & Chute, 2014), and the last major surveys of archival associations and institutions identify email and other e-documents as one of the top three most important collection issues, with the highest marks assigned across all repository types queried (Dooley, 2011; Walch, 2006).

Some recent work has focused on the development of policies, tools, and workflows for preservation of email. For instance, CERP's Email Preservation Parser (2005-2008) and EMCAP focus on email normalization and packaging for preservation. The Persistent Digital Archives and Library System (PeDALS) research project also helps streamline packaging email for preservation. Yet there is need for a tool that supports email appraisal and processing within an integrated solution encompassing all essential components of the digital curation lifecycle — including appraisal, processing, discovery, and delivery. Electronic Archiving System (EAS) designed by Harvard University (which is joining as a Partner for ePADD Phase 2) supports multiple functions across the digital curation lifecycle, though does not currently support pre-acquisition appraisal, accessioning, bulk processing, intellectual arrangement, and discovery. AccessData FTK and ZL Unified Archive can support certain bulk processing functions, though again not intellectual arrangement or online discovery—and furthermore they are proprietary, commercial tools typically unattainable by small and medium-sized institutions. (See **Supporting Document 2: *Comparison of Email Archiving Projects and Software***, for further tool comparison).

We designed ePADD specifically to confront these gaps in the archival management of large volumes of email. Phase 1 of ePADD was based on the MUSE software for exploring personal email archives, discovered during our work as a member of the AIMS project (AIMS, 2012). ePADD Phase 1's upcoming June 2015 release consists of four modules to support archival workflows:

The *Appraisal Module* provides donors, curators, and archivists with a toolset to review and manage an email archive prior to accessioning it to a repository. This module extracts entities (persons, organizations, locations) using a customized NLP toolkit—helping users determine message relevance, identify and flag sensitive messages, and impose access restrictions.

The *Processing Module* is designed for archivists to further perform all functions included in the Appraisal Module, as well as other tasks that prepare the archive for discovery by and delivery to end users, such as reconciling extracted entities with established authority records.

The *Discovery Module* is designed to run under a standalone web server, and allows researchers to browse and search a redacted email collection prior to physically traveling to a repository's reading room to access the full corpus.

The *Delivery Module* provides users with access to the full contents of the processed email archive from a managed workstation in a repository's reading room.

For information about the accomplishments of Phase 1, please see **Supporting Document 3: *Benchmarks achieved in Phase 1***, as well as **Supporting Document 4: *Phase 1 Screenshots***. For information about the technical infrastructure of ePADD Phase 1, please see **Supporting Document 5: *Phase 1 Technical Information***.

ePADD Phase 2 efforts will be directed towards two primary goals:

Broad and Sustained Community Engagement: In Phase 2, we aim to make ePADD a product that is transparent and reflective of the varied needs of many potential stakeholders within the United States and abroad. As such, we will include a diverse set of Partners (institutions providing cost-share) and Collaborators (committed testers not offering cost-share) in the planning and testing process, and will instate a Community Manager to systematically coordinate and oversee such outreach efforts. We intend for our efforts to create and nurture a domestic and international user and developer community that is active and endures beyond the grant period. We will also explore collaboration around archival workflows and processes that support ingest and preservation.

Functional Improvements: First, we aim to significantly expand ePADD's performance capabilities. Certain functions within the Appraisal and Processing modules can be performance-intensive, such as regular expression search, assignment of relevance rankings, and disambiguation of extracted entities. ePADD currently supports performing these functions across archives of up to 250,000 messages on a modest personal computer, particularly important for smaller institutions. In Phase 2 we will optimize ePADD to undertake these functions across archives of up to750,000 messages, while ensuring a satisfactory user experience for all institutions. We have selected 750,000 messages because we believe most personal email archives fall under this category; however, to better meet the need of government and institutional archives which may contain archives in the millions of messages, we will be holding a conference in Y1 for such institutions to inform future software development.

Phase 2 will also advance ePADD's support for restriction and derestriction of materials. Currently, email archives present major challenges with regard to privacy and confidentiality, creating barriers to both donation and usage. Enhanced Phase 2 functionality may include the ability to automate the derestriction of materials flagged by a donor or archivist, eliminating the need for individual management of these messages.

We are further committed to expanding ePADD's scope in several ways. First, we seek to improve ePADD's leveragability and scalability by building in cross-collection discovery capabilities. Similarly, we will explore platforms for cross-institution discovery, as well as public delivery, thereby (a) lowering the participation bar for smaller institutions, and (b) facilitating access to processed email archives approved for public release (e.g. government records requests). Please note that cross-collection and cross-institution discovery will not be impacted by the limit of 750,000 messages set above. We will provide benchmarks on total messages supported in cross-collection and cross-institution discovery upon software release.

Finally, we will implement enhancements to metadata functionalities. Although we will prioritize functional enhancements in consultation with Partners, Collaborators, and the Advisory Board, we are especially committed to incorporating features that expand the ways users can ingest, appraise, discover, and utilize content. Some high-priority enhancements will focus on capturing donor, curator, and end-user supplied metadata (including commentary or analysis that is crowdsourced through the experimental public Discovery module); this will enrich the archives by harnessing user expertise. We will also prioritize expanding the entity types recognized to include fine-grained categories such as books, movies, museums, universities, and companies.

Please see **Supporting Document 6: Development Roadmap** for a more complete list of functional enhancements under consideration for Phase 2.

II. Impact

Satisfying Phase 2 goals will have far-reaching and lasting impact on archival institutions' ability to appraise, process, and provide access to vital cultural heritage materials that otherwise would be unavailable to researchers—filling gaps in our national digital capacity. Phase 2 will also foster a community around open-source programs and methods for bulk archival processing through natural language processing and other techniques, equipping repositories with tools to maintain ePADD's adaptability and contribute to its evolution. Finally, fulfilling Phase 2 goals will lay the groundwork for future efforts applying similar automated workflows and functionalities, including advanced forms of analysis and visualization, to other classes of born-digital materials.

Systematic testing by Partners and Collaborators, the use of survey instruments, and feedback from the community will provide us with reliable information by which to judge ePADD's impact, steer its further development, and achieve the goals specified in Section I: Statement of Need. We have identified five primary targets and related performance indicators relevant to the design, review, and release of Phase 2 deliverables that will help us assess Phase 2 efforts.

- *Performance Scale*: As noted above, we plan to expand ePADD's performance capabilities. To ensure optimal performance, we currently advertise a limit of 250,000 messages per collection. In Phase 2, we are targeting improvement of the program to dramatically increase the message cap to at least 750,000 messages, so that ePADD will be better situated to support both cross-collection and cross-institution discovery of potentially several million messages.
- *Adoption and Community Satisfaction*: We will appraise download numbers through GitHub, perform regular market analysis, and employ survey instruments throughout the grant cycle in order to (1) ascertain the number of institutions that have adopted ePADD, and (2) attempt to gauge the percentage of potential collecting institutions using ePADD for appraisal, processing, discovery, and delivery. The Community Manager will also undertake outreach to scholars and instructors to encourage and then assess ePADD use in the classroom. It is also important that the user community finds the tool responsive, intuitive, and helpful. We will measure community satisfaction with the ePADD software modules through formal and informal surveys. This will include ascertaining whether, on account of ePADD's functionality, potential donors who were previously hesitant to donate email, or institutions that previously did not collect email, now decide to donate or collect. We will also comparatively test the customized NLP to gauge improvements in accuracy, helping to ensure a better user experience.
- *Engagement by Community/Collaborators*: As described in Section V: Communication Plan, we will evaluate community engagement by measuring: unique visits to the ePADD website through Google Analytics; inquiries received over time; quantity of downloads; interest in workshops and presentations; number of followers/feedback on social media; and, bug reporting through YouTrack. We will also measure community engagement by assessing the quantity and quality of community-generated enhancements, including customized lexicons, regular expressions lists, kill-lists of common false positive entities extracted by the NLP

toolkit, and localization files uploaded for community use (collectively, "Community Assets"). The Community Manager will track press and literature coverage, and promote events on the project website ("Project Website") and through social media. We will continue to give presentations, demonstrations, and workshops, and publish papers and other materials to engage with the community and encourage contribution of Community Assets.

- *Engagement by Developers*: Our Advisory Board will help guide our outreach to the developer community and assist with integration into large-scale initiatives (see **Supporting Document 7: *Directory of Advisory Board Members***). We will monitor developer forum postings and usage, including evidence of/interest in crosswalks or other collaborations, software modifications, and the development of localized (non-English) versions of ePADD. To further build developer interest and knowledge around ePADD programming, we will sponsor an ePADD Hackathon in Y3.
- *Engagement by End Users*: The Community Manager will lead efforts to determine scholars' and researchers' usage and opinions of ePADD for working with email archives. Our multi-disciplinary Advisory Board and institutional members/faculty will help steer this assessment. Tracking scholarly and media references to ePADD will also prove useful for such analysis.

Just as ePADD will be unique in its ability to address all components of the digital curation lifecycle for email archives, we also anticipate that ePADD will be notable in its durability and value. Through sustained outreach and commitment to using open-source tools, we aim to build a self-sustaining community. Export functionalities developed through Phase 2 will help ensure that work accomplished through ePADD follows best practices and community standards, and makes progress toward compatibility with existing and future services or technologies. (Please see Section VI: Sustainability for additional information about sustaining project benefits.)

To achieve these lasting outcomes in support of a National Digital Platform, we will produce the following deliverables during Phase 2:

1. ePADD software program: An open-source software package designed by the archival community in collaboration with software programmers. Software and source code will be available for free download through GitHub, and linked through the Project Website. For any crosswalks to other tools or services that are developed, the software and documentation will be linked or hosted on the Project Website.
2. Technical specifications and user documentation will be available through the Project Website and GitHub.
3. Project reports, publications, and presentations (see Section V: Communication Plan).
4. A Project Website hosted through Stanford University Libraries (SUL) (https://library.stanford.edu/projects/epadd). The SUL team will further develop and enhance the existing website as a platform for distributing the ePADD software and documentation and soliciting feedback. The website will also serve as a community hub, offering information about the project, announcements about software developments, links to dedicated qPod forums (http://epadd.nimeyo.com) and to issue reporting/tracking (http://epadd.myjetbrains.com/youtrack/issues/ePADD), and links to Community Assets.

III. Project Design

By ensuring a wide and vibrant user community, conducting stakeholder interviews to guide development of specific functional enhancements, and maintaining iterative development and release cycles, we will achieve the goals and outcomes noted above. Mapping to the project goals and objectives articulated in Section I: Statement of Need, each year of the Phase 2 project prioritizes different build-out, policy, and community development undertakings appropriate to the scope of the project (see Section IV: Project Resources: Personnel, Time, Budget for a detailed timeline; see **Supporting Document 3: *Benchmarks achieved in Phase 1*** for work already accomplished). User testing/reporting will be ongoing throughout the project, and will help with assessing the efficacy of new functionalities.

Year 1 will primarily consist of beginning development work on a priority list of functional enhancements determined through discussion with the Advisory Board and stakeholder interviews with Partners, Collaborators, and other users (e.g. donors, curators, archivists and other cultural heritage professionals, faculty, journalists, and other researchers). We will also build the infrastructure to manage community feedback and Community Assets, as well as provide deliverables. To optimally meet the needs of state and federal repositories and agencies, in Y1 SUL will host a conference geared towards better understanding the requirements of government archives. We will also lay the groundwork for exploring collaborative workflows with other community-supported tools and services that will be pursued during subsequent grant years.

Year 2 will consist of ongoing development work on the priority list of functional enhancements, as determined above and shaped by subsequent discussions with stakeholders (including software and service providers identified in Y1), as well as feedback generated through user testing. In Y2 we will effectuate improvements to restriction management and annotation. We will also investigate and begin implementing a platform that supports the cross-institutional discovery of email archives, as well as public delivery of email archives approved for release either through donor permission, FOIA (or state-equivalent laws), or authorized government entities.

Year 3 will consist of ongoing development work on the functional enhancement priority list (see above), and as guided through subsequent discussions with stakeholders (including other software and service providers identified in Y1/Y2) and feedback generated through user testing. To seed developer interest, in Y3 we will hold an ePADD Hackathon. We will explore preservation policy and implementation issues and solutions, such as exporting processed mail archives to additional formats, exporting metadata to archival and preservation management systems, and consultation and collaboration with entities such as Archivematica around workflows to support preservation. Further, we will develop a framework to export extracted entities as linked open data (LOD). For additional information about the project's commitment to best practices around existing and emerging standards please see Section VI: Sustainability.

To support this project, we have partnered with four organizations, representing a range of public and private academic institutions, as well as cultural and governmental organizations: University of Illinois Urbana-Champaign, Harvard University, University of California, Irvine, and the Metropolitan New York Library Council. Our Partners' key responsibilities will be conducting stakeholder interviews, coordinating user testing and bug reporting, and assisting with prioritizing software functional developments. Additionally, we are approaching a diverse set of institutions to serve as Collaborators. Collaborators have not been asked to contribute cost-share to the project, but have opportunity to beta test the software and provide feedback, as well as help ensure that the project engages with a diverse audience, and that the tool is tested against collections representing a broad diversity of repositories and stakeholder interests.

IV. Project Resources: Personnel, Time, Budget

SUL is providing the time of eight staff members who will balance Phase 2 project responsibilities with existing duties as indicated below. SUL will also offer facilities, equipment and supplies necessary to support the project; please see attached **Budget Form** and **Budget Justification**. Matching funds will be provided as cost-share by staff from SUL and four partnering institutions. Personnel from our partnering institutions, identified below, will benefit from early access to the tool, and the ability to shape its development.

Key Project Staff & Consultants:

Michael A. Keller, Project Director, will be responsible for general supervision of this project, but will not dedicate significant direct effort to necessitate charging his time to the grant. Keller has been SUL University Librarian since 1994, where his responsibilities include oversight and leadership for all project development and community-building activities. As Project Director he is responsible for managing project expenditures.

Glynn Edwards, Head of the Manuscripts Unit and Manager of the Born-Digital Program in Special Collections at SUL, is Project Lead (20% FTE) for ePADD Phase 2. She is responsible for managing project resources and overseeing staff activities. Edwards served as Project Director of the NHPRC-funded ePADD Project. She has built the born-digital program at SUL in collaboration with the Digital Library Systems and Services department (DLSS). Edwards also served as lead archivist on the AIMS Born-Digital Project and was a co-author of the AIMS White Paper.

Peter Chan, Digital Archivist at SUL, is the Project Manager (40% FTE) for ePADD Phase 2. He is responsible for coordination, explanation, review, and communication with the developers and Sudheendra Hangal, the Technical Consultant. He will also monitor and relay feedback gathered through user testing. Chan served as Technical Lead for the NHPRC-funded ePADD Project, and also served as SUL's digital archivist for the AIMS Project. A member of both Special Collections and DLSS, Chan has been at Stanford since 2009 and is a pivotal member of the Born-Digital Program and Lab at SUL. Prior to this Chan was in charge of managing user testing and specification of software development at Bank of America in Hong Kong.

Josh Schneider, Assistant University Archivist at SUL, is the Community Manager (20% FTE) for ePADD Phase 2. Schneider served as UX/UI and documentation lead for the NHPRC-funded ePADD Project. He has presented on ePADD from a policy and technical perspective, and will manage grant reporting, coordinate stakeholder interviews, user testing, and other outreach activities, manage the website and social media, and respond to information requests about the project.

Ixora Technology, will provide the services of a 1) *Software Developer*, 2) *User Interface / User Experience (UI/UX) Designer*, and 3) *Technical Consultant*. Ixora provided these services for ePADD Phase 1, and have extensive experience working with SUL. The *Software Developer*, with a background in machine learning, will build out core machine learning infrastructure and perform general software development for ePADD Phase 2. This individual will also write the technical documentation. The *UI/UX Designer*, with a background in human-computer interface design, will optimize the interface and user experience to ensure broad ePADD adoption. The *Technical Consultant*, Sudheendra Hangal, is the original designer of the MUSE program, the precursor to ePADD. Hangal served as Technical Advisor on ePADD Phase 1. He is Associate Professor of Computer Science, Ashoka University, India. He received his Ph.D. from Stanford University in Computer Science, and has served as Associate Director for Stanford's Mobisocial Lab. Hangal will advise on technical infrastructure, develop text analysis, and review vendor-written code.

Susan Horsfall, SUL Budget Officer, will manage and support personnel with budget oversight. The Office of Sponsored Research will prepare and submit all official grant financial forms.

Partners:

- *Skip Kendall*, Senior Collection Development and Electronic Records Archivist, Harvard University (5% FTE)
- *Margo Padilla*, Strategic Programs Manager, Metropolitan New York Library Council (METRO) (10% FTE)
- *Christopher Prom*, Professor, University Library, Assistant University Archivist, University of Illinois at Urbana-Champaign (5% FTE)
- *Audra Eagle Yun*, Head of Special Collections & University Archives, University of California, Irvine (5% FTE)

Timeline of Specific Activities (includes annual Advisory Board and quarterly Partner Meetings):

Year 1: November 1, 2015 – October 31, 2016

- Nov. 2015: Commit to final infrastructure to support community engagement
- Nov. 2015 – Jan. 2016: Interviews and surveys with Partners, Collaborators, and stakeholders from targeted communities, conducted by Partners and SUL team, to establish list of functional enhancements and assign relative priority
- Jan. 2016: Face-to-Face Partner Meeting
- Mar. – May 2016: Software development on prioritized functional enhancements
- June 2016: Public release of ePADD software (PR1)
- Aug. – Oct. 2016: Software development on aligning with other platforms and services
- Sept. 2016: Conference on ePADD for Government Repositories
- Oct. 2016: Face-to-Face Partner Meeting

Year 2: November 1, 2016 – October 31, 2017

- Nov. 2016: Public release of ePADD software (PR2)
- Jan. – Mar. 2017: Software development on cross-institution discovery and public delivery
- Apr. 2017: Public release of ePADD software (PR3)
- June – Aug. 2017: Software development on restriction and annotation functions
- Aug. 2017: Face-to-Face Partner Meeting
- Sept. 2017: Public release of ePADD software (PR4)

Year 3: November 1, 2017 – October 31, 2018

- Nov. 2017– Jan. 2018: Software development on digital preservation and export formats
- Feb. 2018: Public release of ePADD software (PR5)
- Apr. – June 2018: Software development on issues identified through community outreach
- May 2018: ePADD Hackathon
- July 2018: Public release of ePADD software (PR6)
- Oct. 2018: Final grant report submitted to IMLS

V. Communication Plan

a. *Collaborating with Partners and Reaching Diverse Audiences*: In Phase 2, we will apply coordinated outreach and collaboration efforts to interface with our targeted audience of donors, collecting institutions, end users, and developers. Adoption by a wide variety of repositories is possible, in part, because ePADD is designed to fulfill archival functions associated with email irrespective of institution type or size. To ensure ePADD's functionality has broad appeal, in Phase 2 we are including public and private academic institutions as well as non-profit and governmental institutions among our development Partners. As mentioned previously, we will undertake special outreach to government archives through, among other efforts, a dedicated conference in Y1. In addition to adding these new partnerships, we are also building a diverse pool of Collaborators to serve as the backbone of our user community, perform user testing, and provide feedback on iterative beta releases. Further, computer scientists, journalists, historians, and legal scholars will serve on our Advisory Board, allowing us to maintain presence in an array of communities beyond that of archives and libraries. By incorporating input from our Partners, Collaborators, and Advisory Board we will be better able to serve a variety of end users.

Connecting with our intended audiences requires a harmonized interaction plan between ePADD, Partners, Collaborators, and developers affiliated with other projects supporting a National Digital Platform. We will routinely distribute documentation through a dedicated Project Website hosted by SUL. As detailed in Section II: Impact, we will populate the website with: user and technical documentation; content about the project and latest software releases; links to and information about the software; guidelines to assist integrating foreign language NLP engines; and supporting documentation such as progress reports, results analyses, and stakeholder interviews. We will update user guides after each major release, and will include sections on each module as well as frequently asked questions. The website will also promote and enable collaboration with other ePADD users through Community Asset sharing. We will encourage collecting institutions to provide feedback and help find solutions to common problems through user forums. Software issues, bugs, and new features will be tracked through the open platform YouTrack.

As part of a sustained communication effort, we will deliver training sessions, conference presentations and webinars, and publish project-related papers. To date in Phase 1, we have already spoken at many regional, national, and international conferences, and offered online webinars to inform and educate people about ePADD. In the last year alone, the ePADD team: offered presentations at Society of American Archivists (SAA) 2014, Museums and the Web 2014, Personal Digital Archiving 2014, and CurateCamp 2015; led a workshop at Personal Digital Archiving 2015; and delivered two presentations for the National Digital Stewardship Alliance (2014, 2015). Further, we have provided dozens of demonstrations to individuals and groups in the United States and abroad. We have also drafted professional papers, including on use of ePADD for archival processing in the *Proceedings of Computer Human Interaction* 2015. Through these efforts we have gained traction with repositories, researchers, and journalists—many of whom now widely publicize the ePADD software (Zhang, 2015; Prom, 2014; Hawkins, 2014). Phase 2 will expand along these inroads, and indeed, we already have three upcoming presentations scheduled for RBMS 2015, and SAA 2015. (See **Supporting Document 8: *List of ePADD Presentations and Publications***).

The final major component involves use of community lists and social media platforms. We will share major releases through posts to regional, national, and international listservs, such as for the SAA Metadata and Digital Objects Roundtable, SAA Electronic Records Management Section list, Digital Curation Google Group, American Library Association Digital Curation Interest Group, and ALA-SAA-AAM Combined Committee on Archives, Libraries and Museums. We will continue to publicize ePADD events, presentations, papers, press, and updates through interviews, blogs, and social media (see, e.g., Owens, 2014; Schneider, 2015; and Twitter: @e_padd).

b. *Measuring Engagement and Outcomes*: As also described in Section II: Impact, building community engagement is key to Phase 2's success. Instating a Community Manager will ensure that: 1) communication and outreach efforts are coordinated, 2) the tool is developed and tested by stakeholders that reflect the diversity of potential users, and, 3) ePADD investigates and incorporates crosswalks to common tools and services. We have also strategically invited individuals associated with some of the best-established open-source tools and services, such as Archivematica, Digital Public Library of America (DPLA), and BitCurator, to serve on the Advisory Board, helping us further explore opportunities for collaboration.

To maximize ePADD's adoption in ways that encourage further contributions or refinement by the community, ePADD will remain open-source. As with the Phase 1 release, Phase 2 software and source code will be shared freely online via GitHub. To assess audience engagement and outcomes, we will use Google Analytics to measure visits to our website, as well as click-throughs to the software and source code hosted at GitHub. We will also determine the number of user downloads through GitHub, and extent of bug reporting through YouTrack; these metrics will be made publicly available. We will measure additional engagement by tracking inquiries received over time, interest in workshops and presentations, and amount of followers and feedback on social media.

VI. Sustainability

We will continue to explore sustainability models in consultation with our Advisory Board, Partners, and Collaborators; input from all of these players is critical, as different models may work better for certain types of institutions (non-profit, government, academic, etc.).

Phase 2 build-out of features and services will focus on developing and expanding a practitioner community to ensure ePADD persists in its responsiveness to community needs and practices. To this end, and as discussed above, we will produce and broadly share supporting documentation via the Project Website. We will facilitate user-driven enhancements by encouraging Partners and Collaborators to create and share Community Assets, and support developer creation of custom software modifications. We will utilize public user forums through qPod to support discussion of the software; and, we will make results of stakeholder interviews and surveys, metrics from issue tracking, and ePADD use cases publicly available.

To further ensure ePADD's adaptability within evolving archival standards and practices, Phase 2 will address significant long-term policy and technical issues. ePADD currently supports import of email in MBOX format (considered by many to be a de facto standard) as well as through the IMAP protocol. During Y3 we will explore and test export formats for both the original and processed archives, including MBOX, EML, and XML. To ensure ePADD's extended availability, we will investigate packaging the software as a virtual machine, providing additional precaution against future loss of functionality around processed content. We will develop guidelines around synchronization with tools and services that manage preservation workflows for any export formats selected. We will also support export of message header information and extracted entities in a format facilitating import into NodeXL for network graph analysis and visualization, and export of textual content in a format enabling use of Mallet for topic modeling. Additionally, authority headings assigned to correspondents and extracted entities are currently exportable as CSV files; in Phase 2, we will facilitate export of extracted entities as LOD, enabling the archival and library communities to reuse metadata created through this project.

Together, these actions and build-outs will help ensure that ePADD and the processed email archive (including metadata and other enhancements added by ePADD) are supported beyond the participation of particular developers or institutions, and remain useful in supporting archival workflows and providing access to our vital cultural heritage materials.

Schedule of Completion – Year 1

	2015		2016									
Activity	N	D	J	F	M	A	M	J	J	A	S	O
Advisory Board Meeting	A											
Partner Meeting (Virtual)	P											
Conduct survey and stakeholder interviews with ePADD users	░	░	░									
Partner Meeting (Face-to-Face)			F									
Technical specification and UX/UI design				▒								
Development work on issues identified in user testing and survey					█	█	█					
Partner testing and public release of ePADD software								PR1				
Partner discussion on connection with other platforms and services					░	░	░					
Partner Meeting (Virtual)					P		P					
Technical specification and UX/UI design									▒			
Development work on connection with other platforms and services										█	█	█
Partner testing and public release of ePADD software												
Partner discussion on cross-institution discovery and public delivery										░	░	░
Partner Meeting (Virtual)										P		
Conference on ePADD for representatives of government repositories											C	
Partner Meeting (Face-to-Face)												F

Involves mainly users	░
Involves users and Software Developer	▒
Involves mainly Software Developer	█

PR# = Public Release #
A = Advisory Board Meeting
P = Partner Meeting (Virtual)
F = Partner Meeting (Face-to-Face)

Schedule of Completion – Year 2

	2016		2017									
Activity	N	D	J	F	M	A	M	J	J	A	S	O
Partner testing and public release of ePADD software	PR2											
Technical specification and UX/UI design		▒										
Advisory Board Meeting			A									
Development work on cross-institution discovery and public delivery			█	█	█							
Partner testing and public release of ePADD software						PR3						
Partner discussion on restriction and annotation functions			░	░	░							
Partner Meeting (Virtual)			P		P							
Technical specification and UX/UI design							▒					
Development work on restriction and annotation functions								█	█	█		
Partner testing and public release of ePADD software											PR4	
Partner discussion on digital preservation and export formats								░	░	░		
Partner Meeting (Virtual)								P				
Partner Meeting (Face-to-Face)										F		
Technical specification and UX/UI design												▒

Involves mainly users	░
Involves users and Software Developer	▒
Involves mainly Software Developer	█

PR# = Public Release #
A = Advisory Board Meeting
P = Partner Meeting (Virtual)
F = Partner Meeting (Face-to-Face)

Schedule of Completion – Year 3

	2017		2018									
Activity	N	D	J	F	M	A	M	J	J	A	S	O
Development work on digital preservation and export formats	█	█	█									
Partner testing and public release of ePADD software				PR5								
Partner discussion on previous releases and all user feedback to date	░	░	░									
Partner Meeting (Virtual)	P		P									
Technical specification and UX/UI design					▒							
Advisory Board Meeting						A						
ePADD Hackathon							H					
Development on issues identified through community outreach						█	█	█				
Partner testing and public release of ePADD software									PR6			
Wrap-up										▒	▒	▒
Partner Meeting (Virtual)										P		P
Community building	░	░	░	░	░	░	░	░	░	░	░	░

Involves mainly users	░
Involves users and Software Developer	▒
Involves mainly Software Developer	█

PR# = Public Release #
A = Advisory Board Meeting
P = Partner Meeting (Virtual)
F = Partner Meeting (Face-to-Face)

OMB No. 3137-0071, Exp. Date: 09/30/2015

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:
IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:
If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to complete this form.

Please indicate which of the following digital products you will create or collect during your project. Check all that apply:

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
□	Digital content	Part II
☑	New software tools or applications	Part III
□	A digital research dataset	Part IV

PART I.

A. <u>Copyright and Intellectual Property Rights</u>

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? http://us.creativecommons.org/

Documentation and training materials produced as part of this project will be released under a Creative Commons Attribution 4.0 International License. Software and source code developed by this project will be distributed under an Apache License, Version 2.0 (http://www.apache.org/licenses/). The content of particular email archives may be subject to additional terms and conditions as per the agreements and policies of the host institution.

A.2 What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.

The ePADD software consists of four modules that can be used by a host institution to manage email archives. Additional materials generated by the ePADD project for use by the community (for instance documentation, customized lexicons, extracted entity kill lists, customized regular expression lists—defined collectively in the Narrative as “Community Assets”), or generated by the community and linked or hosted with permission to the ePADD Project Website, will be made available under one of the public licenses described in A.1, and will have no conditions on access or use. Stanford University Libraries makes no claim to ownership over any digital content (e.g. email archives) that may be processed through the ePADD software, or any annotations that may be added by donors, processing archivists, or researchers, for content supplied by non-Stanford entities or institutions.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

ePADD contains various functionalities to facilitate donor and institutional review of materials that may educe privacy concerns (e.g. materials containing Social Security numbers, birthdates, or addresses). In fact, this is one of ePADD’s core strengths to handle and process such content while enabling content restriction. However, all activities relating to privacy, permissions, and rights (including, for example, the restriction of email messages based upon meeting certain criteria) are to be performed subject to the policies, procedures, and risk tolerance of the host institution.

Part II: Projects Creating Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and the quantities of each type and format you will use.

A.2 List the equipment and software that you will use to create the content or the name of the service provider who will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, pixel dimensions).

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during your project and after the grant period.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created during your project (e.g., an Advanced Programming Interface, contributions to the DPLA or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content 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).

D.2 Provide URL(s) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating New Software Tools or Applications

A. General Information

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

Although we predict vast improvements in functionality through Phase 2 software development, we expect the basic functions enabled by each module (Appraisal, Processing, Discovery, and Delivery) will remain largely intact from that of Phase 1:

The Appraisal Module will provide donors, curators, and archivists with a toolset to review and manage an email archive prior to accessioning it to a repository. This module will extract entities using a customized natural language processing (NLP) toolkit—helping to enable users to determine the relevance and importance of email messages, identify and flag sensitive messages, and impose access restrictions. The Processing Module will enable an archivist to perform additional tasks that prepare the archive for discovery and delivery by end users, such as reconciliation of extracted entities with established authority records, as well as export of processed email archives and associated metadata (including extracted entities). The Discovery Module will be designed to run under a standalone web server, and will allow historians, journalists, other professional researchers, and the general public to browse and search a redacted email collection prior to physically traveling to a repository's reading room to access the full corpus. We will specifically look at enabling cross-collection and cross-institution discovery. The Delivery Module will provide historians, journalists, other professional researchers, and the general public with access to the full contents of the processed email archive from a managed workstation in a repository's reading room. We will also explore a public delivery interface for email archives that have been authorized for public release.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

There are many tools that address some piece of appraisal, processing, delivery, and preservation of email archives. AccessData FTK, MailStore, the Collaborative Electronic Records Project, Electronic Archiving System (EAS), Electronic Mail Collaboration Initiative, PeDALS, and ZL Unified Archive all fall into this category. In contrast, ePADD Phase 2 will be an integrated tool to address appraisal, processing, discovery, and delivery. Our use of a Natural Language Processing (NLP) engine to extract entities and reconcile these with authorized headings supports provides functionality unavailable through these other tools. Further, many of the above tools are commercial, potentially rendering them unattainable for archives and repositories of small or moderate size. ePADD Phase 2 will be the only integrated, comparable tool that is open-source, community-developed (thus responsive to varying needs), and freely available to all institutions and the researchers they support.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

Although individual components may change over the course of development to better meet the needs of our users, we expect the basic architecture for ePADD Phase 2 to be similar to that of ePADD Phase 1. Platforms and software used include: Java, Javascript, Apache Tomcat, Java EE Servlet API, Java Mail, Apache Lucene, Apache Tika, the D3-based reusable chart library; Java Application Builder and Launch4J for packaging; other Java libraries from Apache (Lang, commons, CLI, IO, logging, etc.); JSON formatting is performed with the libraries org.json and Gson; a custom NLP; IDEs like IntelliJ Idea and Eclipse; Apache Maven, Ant, and custom shell scripts; Git for source control and YouTrack for issue tracking. We will use the Schema.org vocabulary when publishing our email metadata as linked open data.

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

ePADD Phase 2 will ingest email messages in MBOX format or transferred through IMAP protocol. We expect users may rely on 3rd party software (such as Emailchemy and MailStore) to migrate email messages stored in other formats to MBOX prior to ingest into ePADD. In order to ensure email archives processed using ePADD are exported in a format that is suitable for preservation, we will collaborate with other services that provide open-source digital preservation solutions, and will provide guidelines to our users on exporting files to their preferred solution as appropriate. ePADD currently exports authority records matched to named entities using OCLC FAST in CSV format. In this project we will also investigate exporting and sharing named entities as LOD. In both cases, we expect to adopt workflows to support sharing metadata extracted from email messages using ePADD. We will also support export of metadata and extracted entities to NodeXL and the full text of messages to Mallet.

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

We expect ePADD will continue to be browser-based, and will be compatible with the two most recent versions of Chrome and Firefox. It is currently optimized for Windows 7 and OSX 10.9/10.10 machines, using Java. We will ensure compatibility with additional operating systems (including Windows 10 and upcoming OS releases) in Phase 2. Compatibility with UNIX-based operating systems may be also addressed through use of a virtual machine.

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

We will publish a preliminary development roadmap on the Project Website for discussion with our grant Partners and our Advisory Group. These development goals may be supplemented by additional goals identified during stakeholder interviews and ongoing user testing. The final decision on yearly development goals will be made during our yearly face-to-face Partner meeting, at which the technical consultant will also be present. The development roadmap will be updated accordingly. The software engineer with machine learning experience is responsible for preparing and updating the technical documentation. Issues and bug fixes will be recorded in our public issue-tracking website, YouTrack. Additional discussion will take place in the ePADD qPod forum. Software release notes will accompany major releases shared through GitHub. More comprehensive technical documentation will be published together with the final release through GitHub.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

The Phase 1 public release will be published to GitHub before June 30, 2015. A link to the Phase 1 release will be made available through the ePADD Website (https://library.stanford.edu/projects/epadd/).

C. <u>Access and Use</u>

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

Software and source code developed by this project will be freely distributed under an Apache License, Version 2.0 (http://www.apache.org/licenses/).

C.2 Describe how you will make the software or system available to the public and/or its intended users.

Software and source code developed by this project will be freely distributed via GitHub and linked from the Project Website.

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

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

3. Will you collect any personally identifiable information (PII) about individuals or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release (e.g. data anonymization, suppression of personally identifiable information, synthetic data).

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

6. What documentation will you capture or create along with the dataset(s)? What standards or schema will you use? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

7. What is the plan for archiving, managing, and disseminating data after the completion of research activity?

8. Identify where you will be publicly depositing dataset(s):

Name of repository: ____

URL: ____

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

Supporting Document 1

Works Cited

AIMS Work Group. (2012). AIMS born-digital collections: An inter-institutional model for stewardship. Retrieved from http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf.

Walch, V.I. (2006, Fall/Winter). Archival census & education needs survey in the United States. *The American Archivist*, *69*(2), 291-527.

Brown, M. (2014). Ian McEwan's literary archive bought by Harry Ransom Center for $2m. *The Guardian*. Retrieved from http://www.theguardian.com/books/2014/may/15/ian-mcewan-literary-archive-harry-ransom-center-2m-dollars

Dooley, J.M., & Luce, K. (2010; updated 2011). Taking our pulse: The OCLC research survey of special collections and archives. Dublin, Ohio: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2010/2010-11.pdf

Fung, B. (2015, February 10). Uh-oh: Jeb Bush's 'transparency' effort also exposed Florida residents' personal data. *Washington Post*. Retrieved from http://www.washingtonpost.com/blogs/the-switch/wp/2015/02/10/uh-oh-jeb-bushs-transparency-effort-also-exposed-florida-residents-personal-data/

Kirschenbaum M., et al. (2010). Digital forensics and born-digital content in cultural heritage collections. Council on Library and Information Resources, Washington, D.C. Retrieved from http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf

Hangal, S., et al. (2014). Historical research using email archives in special collections. Proceedings of ACM CHI Conference on Human Factors in Computing Systems. Toronto, Canada. Retrieved from http://mobisocial.stanford.edu/papers/chi2015.pdf

Harry Ransom Center. (2013). McSweeney's archive acquired by Harry Ransom Center. Retrieved from http://www.hrc.utexas.edu/press/releases/2013/mcsweeney/

Hawkins, D. (2014, July/August). Preserving email. *Information Today, 31*(6), 18.

Howard, J. (2011, May 6). On a new frontier for archives, British Library buys poet's 40,000 e-mails. *Chronicle of Higher Education*, *57*(35), A25.

Los Angeles Times. (2011). Sarah Palin emails: the Alaska archive. Retrieved from http://documents.latimes.com/sarah-palin-emails/

Lukesh, S. S. (1999). E-mail and potential loss to future archives and scholarship, or, the dog that didn't bark. *First Monday,* 4(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/692/602

National Archives and Records Administration. (2010). Processing the presidential records of Elena Kagan. Retrieved from http://blogs.archives.gov/aotus/?p=1024

Nicholas, P., & Meckler, L. (2015, March 4). U.S. News: Clinton's email use worries foes of secrecy. *Wall Street Journal*, p. A2.

Noonan, D., & Chute, T. (2014, Spring/Summer). Data curation and the university archives. *The American Archivist*, 77(*1*), 201-240.

Owens, T. (2014, October 20). The ePADD team on processing and accessing email archives. *The Signal*: *Digital Preservation.* Retrieved from http://blogs.loc.gov/digitalpreservation/2014/10/the-epadd-team-on-processing-and-accessing-email-archives/

Pennock, M. (2006, July), Curating e-mails: A life-cycle approach to the management and preservation of e-mail messages. *In DCC Digital Curation Manual*, S. Ross & M. Day (Eds.), Retrieved from http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails

Prom, C. (2014, September 19). Emerging collaborations for accessing and preserving email. *The Signal*: *Digital Preservation*. Retrieved from http://blogs.loc.gov/digitalpreservation/2014/09/emerging-collaborations-for-accessing-and-preserving-email/

Prom, C. (2011, December). Preserving email. *DPC Technology Watch Report 11-01.* Retrieved from http://dx.doi.org/10.7207/twr11-01

Radicati Group Inc. (2015, March). Email statistics report, 2015-2019. Retrieved from http://www.radicati.com/?p=12960

Schneider, J. (2015, February 3). ePADD project makes strides! *Special Collections Unbound*. Retrieved from http://library.stanford.edu/blogs/special-collections-unbound/2015/02/epadd-project-makes-strides

Sinn, D., et al. (2011). Personal records on the web: Who's in charge of archiving, Hotmail or archivists? *Library & Information Science Research*, *33*(2011), 320–330.

Wright, M. (2011, May 10). Why the British Library archived 40,000 emails from poet Wendy Cope. *Wired.* Retrieved from http://www.wired.co.uk/news/archive/2011-05/10/british-library-digital-archives

Zhang, J. (2015). Correspondence as a documentary form, its persistent representation, and email management, preservation, and access. *Records Management Journal*, *25*(1), 78-95.

Email: Process, Appraise, Discover, Deliver – ePADD Phase 2
Stanford University Libraries
June 1, 2015

Supporting Document 2

Comparison of Email Archiving Projects and Software

		SUPPORTED ACTIVITIES											
		Collection Development	Accessioning		Archival Processing				Preservation		Access (end-users)		
		Pre-Acquisition Appraisal	Capture	Normalization	Item-level processing	Bulk processing	Intellectual Arrangement	PII	Packaging	Repository	Online Discovery	Reading Room Delivery	Cost
Open Source Non-Commercial	**CERP** (2005-2008)	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	Nil
	EMCAP	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	
	Archivematica	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	
	PeDALS	🟥	🟨	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	
	ePADD	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	
Proprietary	**EAS**	🟥	🟥	🟨	🟨	🟥	🟥	🟨	🟨	🟨	🟥	🟨	NA
Commercial	**eMailchemy**	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	$
	MailStore	🟥	🟨	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	
	AccessData FTK	🟥	🟥	🟨	🟨	🟨	🟥	🟨	🟥	🟥	🟥	🟨	$$
	ZL Unified Archive	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟥	🟥	🟥	🟨	$$$

🟩 ePADD   🟨 Working Feature   🟥 Not Supported

CERP - Collaborative Electronic Records Project, Smithsonian Institute Archives

EMCAP - e-mail collection and preservation, North Carolina State Archives

Archivematica, Artefactual Systems

PeDALS - Persistent Digital Archives and Library System, Arizona State Library

ePADD - Email: process appraise discover deliver, Stanford University Libraries

EAS - Electronic Archiving System, Harvard University Libraries

eMailchemy, Weird Kid Software LLC

MailStore, MailStore Software GmbH

Access Data FTK, AccessData Group

ZL Unified Archive, ZL Technologies

Adapted from chart developed by Harvard University Libraries

Supporting Document 3

Benchmarks Achieved in ePADD Phase 1

***Appraisal Benchmarks*:** 1) Supports the accessioning of 250,000 messages in a single transfer. 2) Regular expression search and a customizable search lexicon allow the user to flag sensitive content for restriction or non-transfer (including personally identifiable information, personal health information, and other sensitive material that may be governed by FERPA or HIPAA). 3) All messages can be annotated with notes augmenting the content or specifying a restriction period. 4) Bulk review is supported by the inclusion of additional functionalities: correspondent name resolution (merging of correspondents using information extracted from email address fields, which can be corrected by the user), natural language processing (entity extraction and disambiguation for Person, Organization, and Location entities from English text using a customized natural language processing toolkit), image attachment browsing, as well as various other tools to support analysis and visualization of the email archive.

***Processing Benchmarks*:** 1) Partly automates authority work: ePADD compares the entity index compiled by our custom NLP toolkit against popular controlled vocabularies to enable users to link correspondents and extracted entities to established authority records (FAST for subject headings and FreeBase for geographic locations). Headings are checked against DBPedia and displayed according to a relevance algorithm that predicts likelihood of a positive match. The user can manually link records as well. Confirmed headings can be exported as a CSV file for cataloging and creation of finding aids. 2) A redacted version of the email archive (containing only extracted entities) can be exported for discovery through a web-server. 3) A non-redacted version of the email archive can be exported for delivery in the reading room.

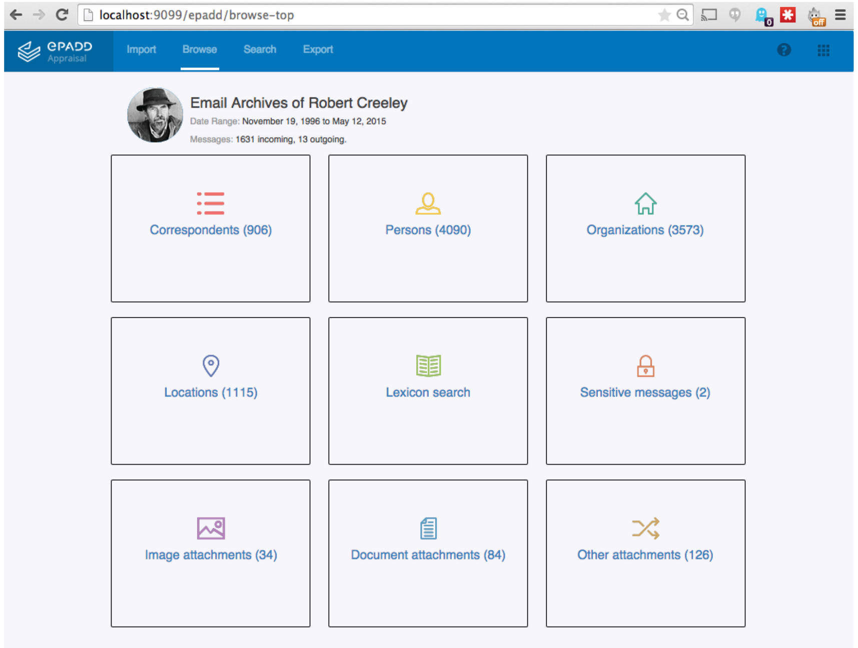
***Discovery Benchmarks*:** 1) A comparative entity search function: ePADD compares and highlights all entities from the email archive that match a given block of text supplied by the user. The user can quickly see which entities from the supplied text appear in the archive, and select them to see the responsive set.

***Delivery Benchmarks*:** 1) Full search, browse, analysis, and visualization functions introduced in previous modules, including image attachment browsing. 2) Ability to flag messages to convey to Public Services for reproduction according to local policies and access provisions.

Email: Process, Appraise, Discover, Deliver – ePADD Phase 2
Stanford University Libraries
June 1, 2015

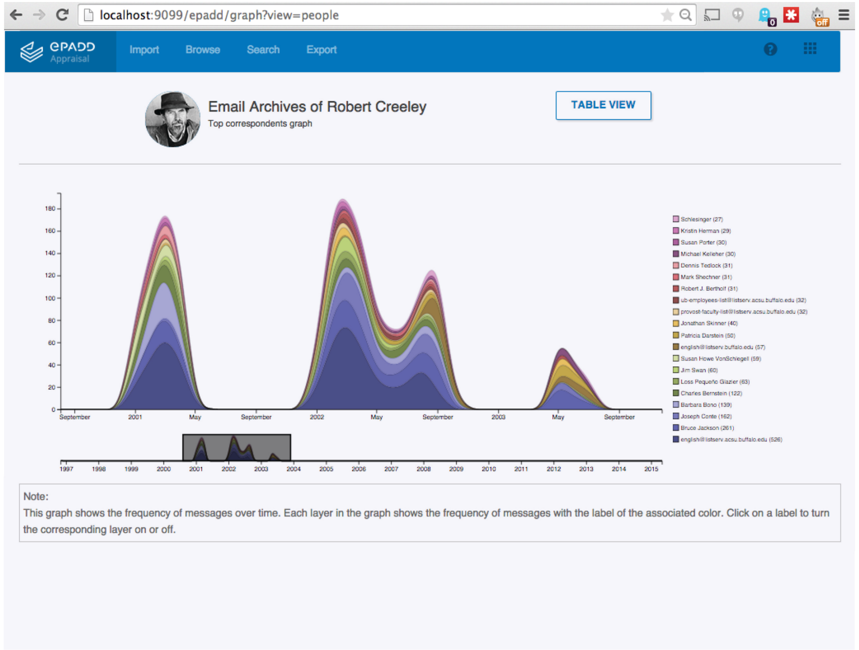
Supporting Document 4

Phase 1 Screenshots



Email Archives of Robert Creeley
Date Range: November 19, 1996 to May 12, 2015
Messages: 1631 incoming, 13 outgoing.

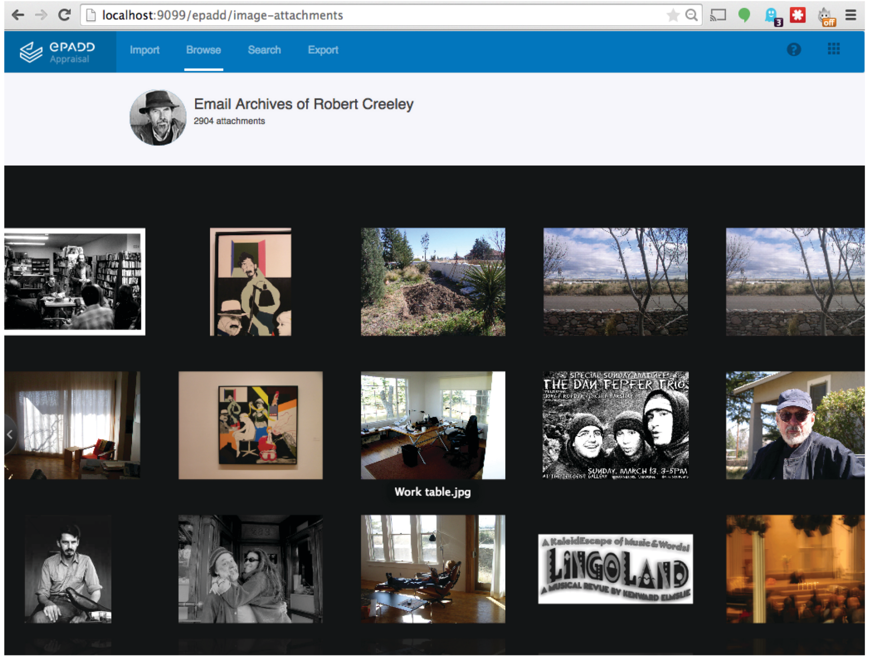
Browse Menu — A hub presenting various functionalities available for the user to review the email archive, including correspondent and extracted entity browsing, a custom lexicon, regular expression search, and attachment browsing.



Email Archives of Robert Creeley
Top correspondents graph

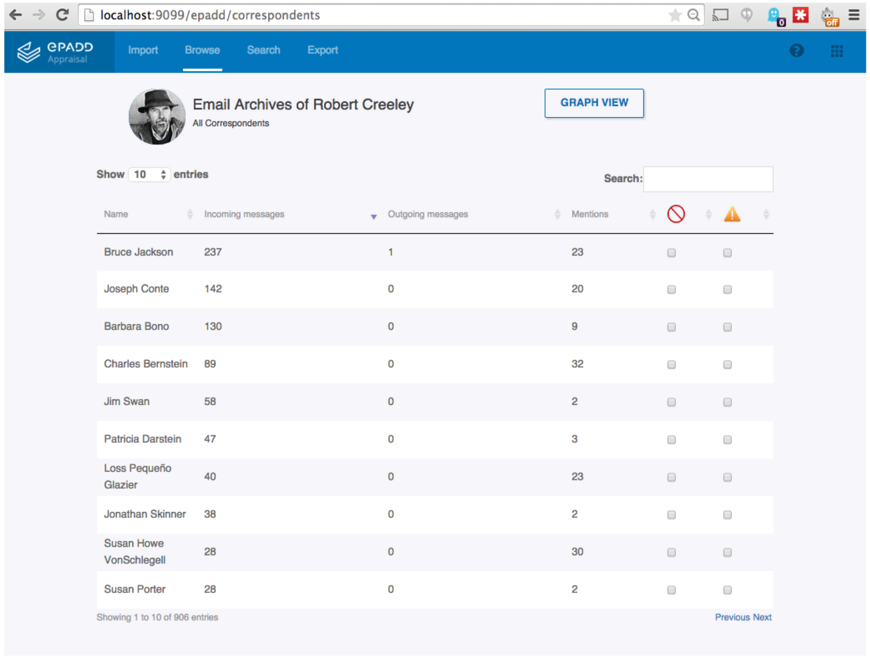
Note:
This graph shows the frequency of messages over time. Each layer in the graph shows the frequency of messages with the label of the associated color. Click on a label to turn the corresponding layer on or off.

Browse Persons Menu — All entities extracted by ePADD can be browsed by entity type in table or graph view. This graph shows the most frequently appearing entities in a selection of poet Robert Creeley's email archive.



Email Archives of Robert Creeley
2904 attachments

Image Attachment Browse Menu — A scrollable image wall. Selecting an image attachment provides a user with information about the associated message as well as the opportunity to directly access that message.



Email Archives of Robert Creeley
All Correspondents

Show 10 entries

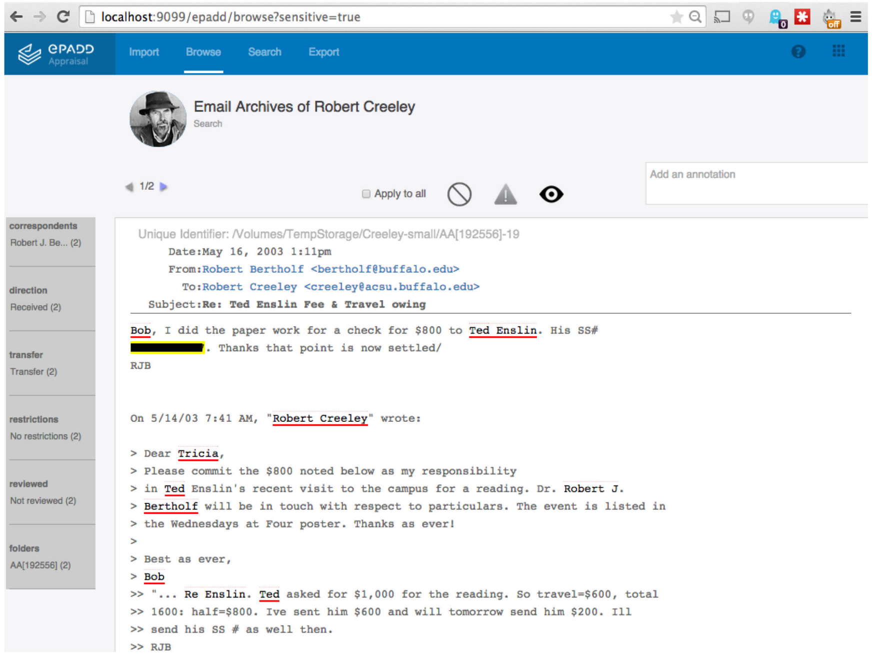
Search:

Name	Incoming messages	Outgoing messages	Mentions	⊘	⚠
Bruce Jackson	237	1	23	□	□
Joseph Conte	142	0	20	□	□
Barbara Bono	130	0	9	□	□
Charles Bernstein	89	0	32	□	□
Jim Swan	58	0	2	□	□
Patricia Darstein	47	0	3	□	□
Loss Pequeño Glazier	40	0	23	□	□
Jonathan Skinner	38	0	2	□	□
Susan Howe VonSchlegell	28	0	30	□	□
Susan Porter	28	0	2	□	□

Showing 1 to 10 of 906 entries

Previous Next

Browse Correspondents Menu — Allows a user to view a table or graph view of the correspondents represented in the collection, and restrict or not transfer messages associated with a particular correspondent.



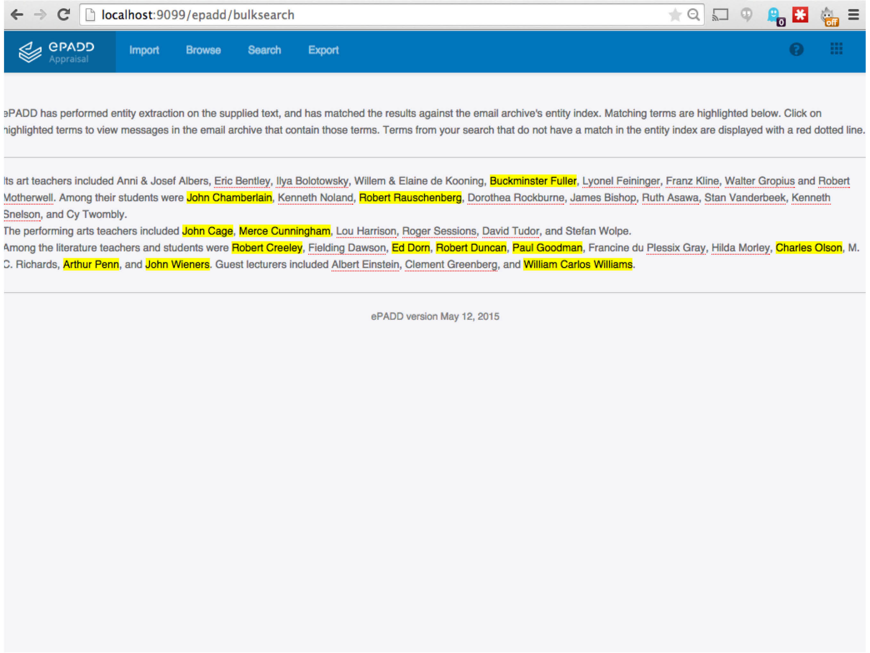
```
Unique Identifier: /Volumes/TempStorage/Creeley-small/AA[192556]-19
     Date:May 16, 2003 1:11pm
     From:Robert Bertholf <bertholf@buffalo.edu>
       To:Robert Creeley <creeley@acsu.buffalo.edu>
  Subject:Re: Ted Enslin Fee & Travel owing

Bob, I did the paper work for a check for $800 to Ted Enslin. His SS#
[redacted]. Thanks that point is now settled/
RJB


On 5/14/03 7:41 AM, "Robert Creeley" wrote:

> Dear Tricia,
> Please commit the $800 noted below as my responsibility
> in Ted Enslin's recent visit to the campus for a reading. Dr. Robert J.
> Bertholf will be in touch with respect to particulars. The event is listed in
> the Wednesdays at Four poster. Thanks as ever!
>
> Best as ever,
> Bob
>> "... Re Enslin. Ted asked for $1,000 for the reading. So travel=$600, total
>> 1600: half=$800. Ive sent him $600 and will tomorrow send him $200. Ill
>> send his SS # as well then.
>> RJB
```

Search Results Menu — ePADD supports search for message subjects, correspondents, extracted entities, the full text of messages and attachments, as well as regular expressions such as social security numbers and bank account numbers. The list of regular expressions is customizable by the user.



ePADD has performed entity extraction on the supplied text, and has matched the results against the email archive's entity index. Matching terms are highlighted below. Click on ighlighted terms to view messages in the email archive that contain those terms. Terms from your search that do not have a match in the entity index are displayed with a red dotted line.

Its art teachers included Anni & Josef Albers, Eric Bentley, Ilya Bolotowsky, Willem & Elaine de Kooning, Buckminster Fuller, Lyonel Feininger, Franz Kline, Walter Gropius and Robert Motherwell. Among their students were John Chamberlain, Kenneth Noland, Robert Rauschenberg, Dorothea Rockburne, James Bishop, Ruth Asawa, Stan Vanderbeek, Kenneth Snelson, and Cy Twombly.
The performing arts teachers included John Cage, Merce Cunningham, Lou Harrison, Roger Sessions, David Tudor, and Stefan Wolpe.
Among the literature teachers and students were Robert Creeley, Fielding Dawson, Ed Dorn, Robert Duncan, Paul Goodman, Francine du Plessix Gray, Hilda Morley, Charles Olson, M. C. Richards, Arthur Penn, and John Wieners. Guest lecturers included Albert Einstein, Clement Greenberg, and William Carlos Williams.

ePADD version May 12, 2015

Bulk Search — ePADD supports a comparative entity search performed between the email archive and a textual passage supplied by the user. Matching entities are highlighted in the search results. Selecting one of the matching entities reveals a menu detailing all of the instances in which the entity occurs in the email achive, and providing direct links to those occurances.

Supporting Document 5

Phase 1 Technical Information

ePADD Phase 1 is written in Java and Javascript and powered by Apache Tomcat (v7.0) using Java EE Servlet API (v3.x) and Java Mail (v1.4.2). Apache Lucene (v4.7) and Apache Tika (v1.8) are utilized for text and metadata extraction, indexing, and retrieval. Charting and visualization is supported using the D3-based reusable chart library (v0.4.10). Java Application Builder and Launch4J (v3.3) are used for packaging on Mac and Windows platforms, respectively. Other Java libraries from Apache (Lang, commons, CLI, IO, logging, etc.) are also used. JSON formatting is performed with the libraries org.json and Gson.

ePADD has implemented its own natural language processing (NLP) toolkit which is used for named entity extraction, disambiguation and other tasks. This toolkit supplants the Apache OpenNLP used in earlier beta versions of the ePADD software, which was itself a replacement for the Stanford NLP used in the original MUSE software on which ePADD is based. We continue to use MUSE as an internal library within ePADD. However, the Apache OpenNLP proved insufficient for needs such as name recognition, and after various rounds of customization, we built our own NLP. This toolkit uses external datasets such as Wikipedia/DBpedia, Freebase, Geonames, and OCLC FAST for LC Subject Headings.

The project is developed with IDEs like IntelliJ Idea and Eclipse, built with Apache Maven, Ant, and custom shell scripts, and tracked using Git for source control and YouTrack for issue tracking.

The ePADD software client is browser-based, and compatible with Chrome 41/42 and Firefox 38/39. It is optimized for Windows 7 and OSX 10.9/10.10 machines, using Java 7/8.

Email: Process, Appraise, Discover, Deliver – ePADD Phase 2
Stanford University Libraries
June 1, 2015

Supporting Document 6

Development Roadmap

Enhance the Natural Language Processing Capability

- Add features to train ePADD on entity extraction, i.e. errors identified by users should update the NLP libraries to prevent similar errors in the future.
- Add the ability to merge several extracted entities into a single entity to support better analysis and visualization.
- Support extraction of additional entities based on the language patterns identified in the initial entity extraction process.
- Analyze email messages by topic (using tools and services such as Mallet and WordNet).
- Extract additional entity types (books, movies, events, etc.) in addition to persons, organizations, and locations.
- Support further refining of extracted entities (e.g. group together "museums" instead of simply "organizations").
- Explore the legal community's use of "predictive coding" employed in electronic discovery during litigation, to better assist patrons in filtering out non-relevant documents.

Enhance the Processing Module Features

- Allow for processing of multiple accessions.
- Allow the processing archivist to redact ONLY the sensitive information in a message (e.g. Social Security numbers) instead of restricting the entire message.
- Enhance the interface for editing and disambiguating correspondents.
- Improve the software to allow for responsive regular expression searches and entity disambiguation for at least 750,000 messages for a single collection.

Enhance the Discovery/ Delivery Module Features

- Develop support for cross-collection discovery/delivery.
- Develop support for cross-institution discovery/delivery.
- Create a public Delivery module to support delivery of email archives without redaction (e.g. email archive released through FOIA requests).
- Allow editing of correspondents, annotation of messages, confirmation of entities, and reconciliation of entities with authority records in the public Delivery module.
- Develop the mechanism to keep and share annotations created by researchers in the Delivery module.
- Allow for institutional branding of the Discovery/Delivery modules.

- Generate a similarity ranking of email messages based on user-selected messages (through the creation of a fuzzy hash).
- Enable the export of email headers and extracted entities to support generation of social network diagrams.

Recommend and Test Preservation Strategy

- Make recommendations on preservation format of email messages (XML, EML, MBOX, etc.).
- Establish a migration process to the recommended preservation format(s).
- Investigate the use of a virtual machine to preserve the ePADD software.
- Investigate the use of a virtual machine and/or "transactional archiving" (http://mementoweb.github.io/SiteStory/) to preserve the ePADD Discovery website.
- Work with common preservation services (such as Archivematica) to facilitate the deposit of email messages, ePADD software, and the ePADD Discovery website to preservation repositories.

Collaboration with other Platforms & Services

- Improve authority record reconciliation using OCLC FAST.
- Investigate authority record reconciliation using other services.
- Improve the entity extraction performed by ePADD by reference to Wikipedia, Geonames, etc.
- Publish extracted entities as linked open data (LOD) to facilitate reuse.
    - Create/adapt vocabulary for extracted data to be available as LOD.
- For websites mentioned in email messages, provide links to the closest version of distributed web archives using the Memento protocol (http://timetravel.mementoweb.org/). If no archived version exists, capture a more recent version of the website.
- Make use of the diseases database maintained by NIH or other sources to highlight messages with potential privacy concerns.
- Investigate the reconciliation of books and periodicals mentioned in email messages to WorldCat.
- Investigate the reconciliation of films mentioned in email messages to IMDb.
- Support export of metadata to Digital Public Library of America (DPLA).
- Support export of headers and entities to NodeXL for network analysis and visualization.

Explore Sustainability Model

- Investigate the policy issues involved in hosting metadata (including extracted entities) for email archives owned by multiple institutions.
- Set up an experimental discovery website to host metadata (including extracted entities) for email archives owned by multiple institutions.

Add Restriction Management/ Annotation Functions

- Allow users to define retention rules, and allow ePADD to inherit previous rule assignments assigned by other clients.
- Allow users to apply retention rules to email messages individually or using bulk application features.
- Implement an automated retention schedule to manage the restriction and de-restriction of materials.
- Allow the user to define actions upon rule conflict.
- Increase support for management of multiple annotation sources (e.g. donor-supplied, archivist-supplied, researcher-supplied).
- Authenticate users in Delivery module in order to add credibility to researcher-supplied annotations.
- Allow users to add multimedia files to email messages as annotations.
- Support crowdsourced transcription/description of files shared through a Public Discovery/Delivery website.

Enhance the Error Handling Capability

- List messages with missing attachments.
- Provide full audit log.
- Provide interface to amend email messages with problems (e.g. insert dates for messages without dates).

Improve User Experience / User Interface

- Engage designer to support improvements to the UX/UI.

Supporting Document 7

ePADD Phase 2: Advisory Board

Sherri Berger is the Product Manager in the Access & Publishing Group, California Digital Library. Berger focuses on helping archives, libraries, and museums expose unique and special collections materials. She provides ongoing product management and communications support to the Online Archive of California and Calisphere services, with an emphasis on envisioning new features to meet user needs. Additionally, Berger currently serves as the Product Manager for the UC Libraries Digital Collection Implementation Project, which will result in a shared, multi-campus digital asset management system and a unified public interface to the UC Libraries' rich and diverse unique resources. She received her MS in Library and Information Science from the University of Illinois at Urbana-Champaign, and her BA in American Studies from Northwestern University.

Andrew Byers is Visiting Assistant Professor in the Department of History at Duke University. His research interests include studies of gender, sexuality, race, class, medical discourse, and the military in the United States in the twentieth century. At Duke, he has also directed a "Humanities Writ Large" laboratory to develop a new undergraduate history curriculum in War, the Military, and Society, a project sponsored by the Andrew W. Mellon Foundation. His next monograph project will continue his explorations of biopolitics in American culture, society, and public policy; the manuscript proposal is currently under consideration by an academic press and the draft will be completed in 2015. He received his PhD and AM in History from Duke, his MA in National Security Studies from Georgetown University, and his BA in History and Political Science from Virginia Tech.

Jackie Dooley is Program Officer at OCLC Research, and an internationally recognized expert in archives and special collections. Within OCLC Research, she leads projects to inform and improve professional archival practice. Her activities have included in-depth surveys of special collections libraries in the U.S./Canada and the U.K./Ireland; development of a series of reports aimed at helping research libraries begin managing their born-digital archival materials; participation in redesign and expansion of Archive Grid; and studying the needs of archival repositories for specialized tools and services.

Marie Hicks is Assistant Professor of History at Illinois Institute of Technology. She is a historian of technology, gender, and modern Europe, specializing in the history of computing. Her recent work focuses on labor and technological change in Britain, and on investigating how 20th century efforts to computerize changed gendered and classed expectations associated with machine work. She studies how collective understandings of social progress are defined by competing discourses of national prestige, labor, and productivity, and how technologies play a formative role in this process. Hicks received her PhD and MA from Duke University, and her AB from Harvard University.

Jeremy Leighton John is Curator of e-manuscripts at the British Library and Principal Investigator of the Digital Lives Research Project. John has been Curator of eMANUSCRIPTS and Scientific Curator in the Department of Western Manuscripts at the British Library since 2003, having been Specialist Curator for the W. D. Hamilton Archive from 2000. Previously he worked as a cataloguer of bioacoustic collections in the British Library Sound Archive. He served as Principal Investigator for Digital Lives, a major research project focusing on personal digital archives and their relationship with research repositories. He is a member of the Library Committee of the Royal Society, and of the Advisory Committee of the National Cataloguing Unit for the Archives of Contemporary Scientists. He is also a Fellow of the Linnean Society of London and of the Royal Geographical Society, and is a member of the British Society for the History of Science. During his career, he has won several scholarships and prizes, including one for writing, has conducted both field and theoretical research, and has given talks and published articles on scientific, archival and historical topics, in both scholarly and popular forms. Further, he has been working with hybrid (digital and analogue) collections of living as well as deceased scientists, and has also been adapting technologies and procedures for forensically capturing, authenticating and making available the digital equivalent of analogue personal archives and manuscripts. John has a DPhil in Zoology from Merton College, Oxford.

Monica Lam is a Professor in the Computer Science Department at Stanford University, which she joined in 1988. She received a B.Sc. from University of British Columbia in 1980 and a Ph.D. in Computer Science from Carnegie Mellon University in 1987. She is the Faculty Director of the Stanford MobiSocial Computing Laboratory and a co-PI in the POMI (Programmable Open Mobile Internet) 2020 project, which is an NSF Expedition started in 2008. Her current research interests lie in building an open and federated social computing infrastructure. She has worked in the areas of compiler optimization and software analysis to improve security.

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches courses on archival administration; records management; digital curation; understanding information technology for managing digital collections; and acquiring information from digital storage media. He is a lead organizer and instructor for the DigCCurr Professional Institute, a week-long continuing education workshop on digital curation, and he teaches professional workshops on the application of digital forensics methods and principles to digital acquisitions. Lee's primary area of research is the long-term curation of digital collections. He developed "A Framework for Contextual Information in Digital Collections," and edited and provided several chapters to *I, Digital: Personal Collections in the Digital Era* published by the Society of American Archivists. Lee is Principal Investigator of the BitCurator project, which is developing and disseminating open-source digital forensics tools for use by archivists and librarians.

Evelyn McLellan is President of Artefactual, developers of Archivematica and AtoM. She is responsible for directing Artefactual's business operations and strategy. She also works as a senior systems analyst on Artefactual's development and client projects. Prior to joining Artefactual, McLellan had over 10 years experience as an archivist and records manager at a number of organizations including the City of Vancouver Archives and the Insurance Corporation of British Columbia. She has worked as co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project and as Adjunct Professor at the

University of British Columbia's School of Library, Archival, and Information Studies from 2004-2010. She received her MA in Archival Studies from University of British Columbia in 1997.

Philip Malone is Professor of Law, and Inaugural Director, Juelsgaard Intellectual Property and Innovation Clinic at Stanford University. He is a leading expert in IP, innovation and cyberlaw, and brings to the position nearly a decade of experience in clinical education and another 20 years of antitrust and technology litigation, including work for the Department of Justice. His clinical work and scholarship is focused on understanding and promoting sound innovation and exploring how intellectual property and competition policy in high-tech industries affect it. His work also looks at ways in which to encourage broad opportunities for creativity, online expression, open access and dissemination of information, and increased access to justice. His teaching has addressed the relationship between legal policy and innovation, including the role of competition and antitrust law, intellectual property, privacy, and security law.

Mark Matienzo is the Director of Technology for the Digital Public Library of America. Prior to joining DPLA, Matienzo worked as an archivist and technologist specializing in born-digital materials and metadata management, at institutions including the Yale University Library, The New York Public Library, and the American Institute of Physics. Matienzo received a MSI from the University of Michigan School of Information and a BA in Philosophy from the College of Wooster, and was the first awardee (2012) of the Emerging Leader Award of the Society of American Archivists.

T. Christian Miller is an award-winning investigative reporter, author, and war correspondent for ProPublica, an independent non-profit newsroom that produces investigative journalism in the public interest. His investigative reports have focused on how multinational corporations operate in foreign countries, documenting human rights and environmental abuses. He has also covered wars and U.S. campaigns. Miller is a pioneer in the field of computer-assisted reporting, and was awarded a Knight Fellowship at Stanford University in 2012 to study innovation in journalism.

Jessica Moran is Assistant Digital Archivist at the Alexander Turnbull Library, National Library of New Zealand where she is responsible for supporting the acquisition, ingest, and management of born-digital heritage collections. She works closely with the National Digital Heritage Archive program, contributing to digital preservation policy and planning as well as system testing and requirements analysis. She has previously worked in university, special, and government libraries and archives, most recently as an archivist at the California State Archives where she worked in the electronic and legislative records programs. She has a BA from UC Berkeley, an MLIS with a concentration in Archives from San Jose State University, and an MA in History from San Francisco State University.

David Rosenthal is founder of the LOCKSS Program, aimed at long-term preservation of web-published materials (ejournals, books, blogs, websites, archival materials, etc). He built and tested the initial prototype, developed the OpenBSD-based network appliance technology that LOCKSS peers used for the first 5 years of production, and was part of the research team that developed the award-winning fault- and attack-resistant peer-to-peer network technology that underlies the LOCKSS network. He currently works on economic models for long-term storage. David received an MA degree from Trinity College, Cambridge and a PhD from Imperial College, London. He is the author of several technical publications and holds 23 patents.

Ben Shneiderman is Professor in the Department of Computer Science and Founding Director (1983-2000) of the Human-Computer Interaction Laboratory at the University of Maryland. He was elected as a Fellow of the Association for Computing (ACM) in 1997 and a Fellow of the American Association for the Advancement of Science (AAAS) in 2001. He received the ACM SIGCHI Lifetime Achievement Award in 2001. Shneiderman is the co-author, with Catherine Plaisant, of *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (5th ed., April 2009).

Marc A. Smith is a sociologist specializing in the social organization of online communities and computer-mediated interaction. Smith leads the Connected Action consulting group and lives and works in Silicon Valley, California. Connected Action applies social science methods in general and social network analysis techniques in particular to enterprise and Internet social media usage. He is the co-editor, with Peter Kollock, of *Communities in Cyberspace* (Routledge), a collection of essays exploring the ways identity, interaction, and social order develop in online groups. Smith received a BS in International Area Studies from Drexel University in Philadelphia in 1988, an MPhil in social theory from Cambridge University in 1990, and a Ph.D. in Sociology from UCLA in 2001. He is an affiliate faculty at the Department of Sociology at the University of Washington and the College of Information Studies at the University of Maryland. Smith is a Distinguished Visiting Scholar at the Stanford University Media-X program.

Kam Woods is a Research Scientist in the School of Information and Library Science at the University of North Carolina at Chapel Hill. He is currently Technical Lead on the BitCurator project, and works with Cal Lee developing techniques and tools to assist in long-term archiving of born-digital data. Woods' research focuses on long-term preservation of born-digital materials. He is interested in interdisciplinary approaches that combine technologies and expertise in the areas of archiving, computer science, and digital forensics for the purpose of enabling and maintaining access to digital objects that are at risk due to obsolescence. Prior to his current work at UNC, Woods worked with Lee on the development of educational materials to support the use of realistic forensic datasets in professional training and to identify and explore novel uses of forensic data and tools in the context of digital archives.

Supporting Document 8

List of ePADD Presentations and Publications

Presentations and Workshops

Edwards, G. (Forthcoming 2015). ePADD and automated processing. Society of American Archivists (SAA) Annual Meeting, Cleveland, OH.

Schneider, J. (Forthcoming 2015). Out of the frying pan and into the reading room: Approaches to serving electronic records. SAA Annual Meeting, Cleveland, OH.

Edwards, G., & Chan, P. (Forthcoming 2015). Automated email processing using ePADD. ALA RBMS Preconference, Oakland, CA.

Edwards, G., Chan, P., & Schneider, J. (Forthcoming June 2015). ePADD. Archiving Email Symposium sponsored by Library of Congress and the National Archives and Records Administration, Washington, D.C.

Chan, P. (2015). Email – process, appraise, discover, deliver. CurateCamp on Born-digital Workflow, New York, NY.

Edwards, G., & Schneider, J. (2014). ePADD: Live demonstration of appraisal and processing modules. Email Interest Group, National Digital Stewardship Alliance.

Schneider, J. (2014). ePADD: project overview and recent developments. Manuscript Repositories Section, SAA Annual Meeting, Washington, D.C.

Schneider, J. (2014). Overview of the ePADD appraisal module. Acquisitions and Appraisal Section, SAA Annual Meeting, Washington, D.C.

Edwards, G., & Chan, P. (2014). ePADD—Assessing tools and best practices for email preservation and access in art museums. Museum and the Web Deep Dive, Baltimore, MD.

Chan, P. (2014). ePADD. Personal Digital Archiving Conference. Indianapolis, IN.

Hangal, S. (2013). Providing access to email archives for historical research. Personal Digital Archiving Conference, College Park, MD.

Hangal, S. (2012). Muse: A tool to mine and visualize large email archives. Investigative Reporters and Editors Conference, Boston, MA.

Edwards, G. (2012). The enigma of email archives: How to process and provide discovery environments. ALA RBMS Preconference, San Diego, CA.

Chan, P. (2012). Processing and delivering email archives in special collections using MUSE. Personal Digital Archiving Conference, San Francisco, CA.

Hangal, S. (2012). Putting personal archives to work: reminiscence, search and browsing. Personal Digital Archiving Conference, San Francisco, CA.

Hangal, S., et al. (2012). Processing email archives in special collections. Digital Humanities Conference, Hamburg, Germany.

Publications by ePADD Team

Hangal, S., et al. (2014). Historical research using email archives in special collections. Proceedings of ACM CHI Conference on Human Factors in Computing Systems. Toronto, Canada.

Chan, P., et al. (2013). New horizons in personal archiving: 1 Second Everyday, myKive and MUSE. In *Personal Archiving: Preserving Our Digital Heritage*, Donald T. Hawkins (ed.), Medford, NJ: Information Today, Inc.

Publications by Others

Hawkins, D. (2014, July/August). Preserving email. *Information Today, 31*(6), 18.

Owens, T. (2014, October 20). The ePADD team on processing and accessing email archives, *The Signal*: *Digital Preservation.* Retrieved from http://blogs.loc.gov/digitalpreservation/2014/10/the-epadd-team-on-processing-and-accessing-email-archives/

Prom, C. (2014, September 19). Emerging collaborations for accessing and preserving email, *The Signal*: *Digital Preservation*. Retrieved from http://blogs.loc.gov/digitalpreservation/2014/09/emerging-collaborations-for-accessing-and-preserving-email/

Prom, C. (2011, December). Preserving email. *DPC Technology Watch Report 11-01.* Retrieved from http://dx.doi.org/10.7207/twr11-01

Zhang, J. (2015). Correspondence as a documentary form, its persistent representation, and email management, preservation, and access. *Records Management Journal*, *25*(1), 78-95.

Original Preliminary Proposal

Proposal Overview

ePADD is a software package that supports archival processes around the ingest, appraisal, processing, discovery, and delivery of email archives. During the past eighteen months (Phase 1), relying in part upon NHPRC funding provided through June 2015, we have proven the concept of using natural language processing, automated metadata extraction, and other batch processes to support archival workflows and to provide access to otherwise hidden cultural heritage materials. With this grant application for Phase 2 (projected for October 2015 – September 2018), we seek to move ePADD to a web-based service and greatly expand the program's functionality, including build-out of feature requests identified in Phase 1. We also seek to ensure broad adoption of ePADD through stakeholder interviews, expanded user testing, UI enhancements, community engagement, and partnerships.

Over the course of ePADD's development, need and demand for a tool of this potential scope have continued to build. Email archives present a singular window into contemporary history—one that archival organizations desire to make available for research, but for which they face difficulties due to the screening, processing, and access challenges of the medium and the sheer volume of material. ePADD is designed specifically to address these challenges for large volumes of email, and in a manner that is both customizable and scalable. In this way, ePADD has the potential to be a truly transformative tool to support the appraisal, processing, discovery, and use of rich digital content, and lays the groundwork for future efforts applying similar workflows and functionalities to other classes of born-digital materials.

Proposed Work Plan

Under the direction of Stanford University Libraries, in collaboration with our partners indicated below, and with counsel from an Advisory Board, we will apply funds to four areas:

(1) *Specification Development*. Phase 1 work with collaborating cultural and academic institutions and the Advisory Board helped identify more than two-dozen features critical for maximizing ePADD's adoption and usefulness. In Phase 2, a project manager / community manager will coordinate stakeholder interviews with collaborating institutions to inform software development. Technical advising by Sudheendra Hangal (the developer of Muse, the precursor to ePADD) will continue.

(2) *Software Development*. A software developer will implement the components specified, including expanding use and accuracy of natural language processing (NLP) to enable more advanced entity extraction and resolution, greater support for bulk processing and content analysis, content redaction, support for creating local authorities and exporting linked data, and support for preservation.

(3) *User Testing*. Phase 1 will culminate (in April 2015) in the release of an open-source tool from which extensive user feedback will be generated. This feedback will form the basis of a comprehensive user-testing program for Phase 2, which will be managed by the project manager / community manager.

(4) *Community Engagement*. In collaboration with partner universities (e.g. Harvard University, University of Illinois, New York University) and government and cultural institutions (e.g. Library of Congress, Metropolitan New York Library Council), the project manager / community manager will assume responsibility for community engagement, including developing and managing the user community, coordinating communications and activities, and creating supporting documentation; inclusion of such a role is important at this juncture to ensure that the Phase 1 prototype becomes useful for a wide variety of user communities.

Furtherance of National Digital Infrastructure Funding Priorities

There are few current tools to support appraisal, processing, discovery, or delivery of email, and those that exist are insufficient. The upcoming April 2015 release of ePADD provides the profession with implements to begin addressing these challenges. Phase 2 development through IMLS NLG funding would ensure that ePADD endures in improving the discoverability and research utility of this digital content, and will continue to build an infrastructure to support the long-term stewardship and provision of meaningful access to these important materials. Email archives including metadata and extracted entities will be exported for discovery via a public website to allow for cross-institution searching. We will also explore moving ePADD software to a web service; unified access to the web service would enable the inclusion of a machine learning-based training module for the NLP Toolkit, incrementally improving the accuracy of ePADD automatically as it becomes more widely used.

Performance Goals & Projected Impact

ePADD's Phase 2 performance goals stress expanding content, use, and preservation functionalities, and cultivating broad community engagement at all stages of tool development.

Content, Use, and Preservation Goals. ePADD currently supports archives of up to 250,000 messages; Phase 2 projects a build-out of capabilities to support up to 750,000 messages, better aligned to real world conditions as reported by current collaborators. New features will expand the range of ways users can ingest, appraise, discover, and utilize content. Phase 2 will enable the export of both the original and processed email archive in XML and MBOX formats, as well as the export of all headings and extracted entities, which will be published as linked open data. The final release of ePADD during this grant period will be packaged as a virtual machine. Together, these build-outs will help ensure that the email archive and enhancements added by ePADD remain useable irrespective of developer support or modifications to the platform.

Tool Development and Community Engagement Goals. Phase 2 will improve functionality around donor, curator, and end-user supplied metadata, enriching the archive by providing the mechanism to harness this expertise, and will also offer superior tools for restriction and de-restriction of materials. Currently, email archives present major challenges with regard to privacy and confidentiality, which create barriers to both donation of materials as well as their use. Enhanced Phase 2 functionality will include the ability to redact sensitive information, increasing access to messages that would otherwise be restricted. NLP tools will continue to be refined to increase accuracy. In order to maximize the adoption of the tool in a way that would encourage further work by the community, ePADD will remain open-source. The software will be shared freely online, and the source code will be made available on GitHub. Supporting documentation will be produced and broadly shared. Instating a Phase 2 project manager / community manager will ensure that outreach efforts are coordinated, that the tool is being developed and tested by stakeholders that reflect the diversity of potential users, and that ePADD incorporates crosswalks to the most common tools and services.

Estimated Budget

Programmer and consulting fees:	$530,000.
In-person meetings and conference travel:	$60,000
Equipment and service costs:	$20,000
Total requested:	$610,000