Using ePADD to Appraise, Process, and Provide Access to Historically and Culturally Valuable Email Collections

Tutorial Abstract

J. Schneider
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
josh.schneider@stanford.edu

ABSTRACT

This tutorial will provide participants with the knowledge and experience to begin to implement ePADD at their institution in order to appraise, process, and provide access to email archives in their collections. There will be a brief lecture and discussion that focuses on the motivation for development of the software, and a comparison of the software in the context of other tools and services. Several technical concepts will be introduced. The remainder of the tutorial will be devoted to demonstration and hands-on exercises taking users through specific modules and functionalities.

Participants will learn how to ingest email in ePADD from MBOX files or through IMAP, the processes ePADD carries out upon ingest (de-duplication, name resolution, named entity recognition/ classification), how to instruct a donor on appraising and transferring their own email, how to use browse and search features to screen for and restrict confidential or sensitive messages and attachments, how to annotate messages, how to assign authority records to correspondents using ePADD, and how to export mail between ePADD modules. Participants will also learn how researchers can use these same tools within ePADD to discover, browse, and search relevant messages, including how they can review and request specific messages and attachments.

The tutorial will also include discussion on overcoming potential implementation challenges, as well as opportunities to participate in ePADD's development. The tutorial will be beneficial for all those with a responsibility or interest in processing and/or providing access to email archives.

CONCEPTS

• **Computing Methodologies** → **Artificial Intelligence**; *Natural language processing* • **Computing Methodologies** → **Machine Learning** • **Information Systems** → **World Wide Web**; *Web applications;* Internet communications tools; *Email*

KEYWORDS

Acquisition, Archival appraisal, Archival processing, Archives, Descriptive metadata, Email, Named entity recognition, Natural language processing, Privacy, Redaction, Screening, Web access

1 ePADD SOFTWARE

ePADD is free and open-source computational analysis software that allows individuals and institutions to appraise, process, and make discoverable and fully accessible for research email of potential historical or cultural value. [2] The software primarily accomplishes this goal by incorporating techniques from computer science and computational linguistics, including natural language processing, named entity recognition, and other statistical machine learning-associated processes [1].

2 ePADD SOFTWARE ARCHITECTURE

2.1 Appraisal

Appraisal provides donors, curators, and archivists with a toolset to review and manage an email archive prior to accessioning it to a repository. ePADD can gather email from multiple sources. Upon ingest, ePADD de-duplicates messages, resolves correspondent names from the address book, and extracts fine-grained entities using a custom NLP toolkit. These functionalities and others enable users to determine the relevance and importance of email messages, identify and flag confidential, restricted, or legally-protected information, and impose access restrictions prior to transfer.

2.2 Processing

Processing is designed for an archivist to further perform all functions included in the Appraisal module, including scanning for confidential, restricted, or legally-protected information, as well as other tasks that prepare the archive for discovery by and delivery to end users, such as reconciliation of correspondents and extracted entities with established authority records.

2.3 Discovery

Discovery 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. Only metadata from the processed email archive is published online.

2.4 Delivery

Delivery provides users with access to the full contents of the unrestricted portions of a processed email archive, including attachments, from a managed workstation in a repository's reading room.

3 TUTORIAL FORMAT

There will be a brief lecture and discussion that focuses on the motivation for development of the software, and a comparison of the software in the context of other tools and services. Several technical concepts will be introduced. The remainder of the tutorial will be devoted to demonstration and hands-on exercises taking users through specific modules and functionalities, as well as a brief discussion on overcoming potential implementation challenges, as well as opportunities to participate in ePADD's development.

4 INTENDED AUDIENCE

The tutorial will be beneficial for all those with a responsibility or interest in appraising, processing, and/or providing access to email archives, or who are interested in familiarizing themselves with the technologies utilized to accomplish these goals, in order to potentially apply them to other file genres or formats.

5 EXPECTED LEARNING OUTCOMES

This tutorial will provide participants with the knowledge and experience to begin to implement the free and open source software ePADD at their institution in order to appraise, process, and provide access to email archives in their collections.

Participants will learn how to ingest email in ePADD from MBOX files or through IMAP, the processes ePADD carries out upon ingest (de-duplication, name resolution, named entity recognition/ classification), how to instruct a donor on appraising and transferring their own email, how to use browse and search features to screen for and restrict confidential or sensitive messages and attachments, how to annotate messages, how to assign authority records to correspondents using ePADD, and how to export mail between ePADD modules. Participants will also learn how researchers can use these same tools within ePADD to discover, browse, and search relevant messages, including how they can review and request specific messages and attachments.

Participants will also become aware of the resources that are available for learning more about the software and overcoming potential implementation challenges, as well as opportunities to participate to participate in the user community and join in ePADD's development.

6 INSTRUCTOR BIOGRAPHY

Josh Schneider is Assistant University Archivist at Stanford University, where he acquires and supports researcher use of Stanford University records, faculty papers, and materials documenting campus and student life. His case study on appraisal of electronic records appeared in the latest volume of the Society of American Archivists' *Trends in Archival Practice* series. Josh is also Community Manager for ePADD, an open-source software package that uses named entity recognition and other NLP-driven processes to support the appraisal, processing, discovery, and delivery of email archives. Josh serves on the editorial boards of The American Archivist, Journal of Western Archives, and the blog of SAA's Electronic Records Section (BloggERS!). He received an MLIS from Simmons College and a BA in Philosophy from Brown University.

ACKNOWLEDGMENTS

ePADD development is managed by Stanford University's Department of Special Collections & University Archives, part of Stanford University Libraries [3]. The ePADD development team is composed of Glynn Edwards, Peter Chan, Josh Schneider, and Sudheendra Hangal [4]. Development work is carried out in collaboration with partners at Harvard University, the Metropolitan New York Library Council (METRO), University of Illinois at Urbana-Champaign, and University of California, Irvine. Funding for current ePADD development is provided through an Institute of Museum & Library Studies (IMLS) National Leadership Grant (NLG) for Libraries. Development for the initial 2015 release of ePADD was primarily funded by the National Historical Publications and Records Commission (NHPRC).

REFERENCES

[1] Email: Process, Appraise, Discover, Deliver -- ePADD Phase 2. *Project Proposal*, National Leadership Grant for Libraries. Retrieved June 27, 2017, from Institute of Museum and Library Services: https://www.imls.gov/grants/awarded/lg-70-15-0242-15

[2] ePADD repository, 2017. Retrieved June 27, 2017, from Github: https://github.com/EPADD/epadd

[3] ePADD software, 2017. Retrieved June 27, 2017, from Stanford University Libraries: http://library.stanford.edu/projects/epadd

[4] 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 June 27, 2017, from Stanford University: https://mobisocial.stanford.edu/papers/chi2015.pdf