ePADD: Supporting Archival Appraisal, Processing, and Research for E-mail Collections

Josh Schneider
Stanford University, josh.schneider@stanford.edu

Follow this and additional works at: https://lib.dr.iastate.edu/macnewsletter
Part of the Archival Science Commons

Recommended Citation
Schneider, Josh (2016) "ePADD: Supporting Archival Appraisal, Processing, and Research for E-mail Collections," MAC Newsletter: Vol. 43 : No. 3 , Article 8.
Available at: https://lib.dr.iastate.edu/macnewsletter/vol43/iss3/8

This Electronic Currents is brought to you for free and open access by the Journals at Iowa State University Digital Repository. It has been accepted for inclusion in MAC Newsletter by an authorized editor of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
ePADD is an open-source downloadable software package developed by Stanford University’s Special Collections and University Archives that harnesses and develops natural language processing (NLP) and other computational analytic methods to automate the process of reviewing e-mail archives and providing access to them. The public release of ePADD was made available through Github on June 30, 2015, following two years of development funded through the National Historical Publications and Records Commission.

A three-year IMLS grant beginning in November 2015 is funding a second phase of development for ePADD. This phase will focus on building out additional functionality that advances the formation of a National Digital Platform through expanding the program’s scalability, usability, and feature set. Special Collections and University Archives will undertake this work with partners at the University of California, Irvine; the University of Illinois Urbana-Champaign; Harvard University; and the Metropolitan New York Library Council (METRO).

Overview
E-mail has become a dominant instrument of modern communication. The archival e-mail collections of recent authors and public figures can provide unique windows into contemporary society, helping scholars uncover cultural trends and insight into those individuals’ self-expressions, collaborations, networks, and transactions. To date, however, it has been difficult for archival institutions to collect, process, and provide access to these e-mail archives due to concerns about privacy as well as the difficulty of processing large archives containing potentially hundreds of thousands of messages gathered over many decades.

ePADD provides donors, curators, and archivists with tools to help automate the process of screening these archives for confidential information and prepare them for analysis by scholars. Using the same tools on the public-facing side, ePADD supports browsing, searching, and visualization of e-mail archives, extending the ability of scholars to analyze these materials.

Underlying Functionality
ePADD relies in part on machine learning functionalities aligned with developments in natural language processing, which allow for the automation of many of the processes associated with traditional archival processing and research. One of the primary functionalities built into the first release of ePADD is the custom named entity recognizer (NER) designed to recognize people, organizations, and locations. Incorporation of the NER allows a user to browse and visualize an e-mail corpus by person, organization, or location. The NER also drives a comparative “query generator,” which permits the entity index for a currently browsed collection to be compared against the entity index of another text-based corpus such as the full text of a Wikipedia page, dissertation, or monograph.

Additionally, ePADD supports automated disambiguation and resolution of correspondents and named entities identified in the corpus to help indicate whether the Robert mentioned in one e-mail message is the same Robert mentioned in another. ePADD also supports reconciliation of correspondents and named entities with authorized headings using controlled vocabularies such as Library of Congress Subject Headings (LCSH) provided through OCLC Faceted Application of Subject Terminology (FAST).

In the case of ambiguous matches, ePADD indicates relative confidence levels in a given match by...
comparing the extracted entity index for all messages within the e-mail archive containing the named entity in question with the DBpedia entry for a given entity. A high relative confidence level indicates a higher probability that the entities mentioned are the same. The ePADD NER takes advantage of the e-mail address book to better identify and disambiguate correspondents and weigh them more heavily in this ranking process.

Screening E-mail Collections for Confidential Information

One major challenge to repositories in making e-mail collections available for researchers is the difficulty of manually reviewing potentially hundreds of thousands of messages to screen for privacy issues and confidentiality concerns. For this reason, many archives are simply unable to open e-mail collections for research. In addition to integrating NLP functionality to support the donor, curator, or archivist in preparing the e-mail archive for use by researchers, ePADD supports automated screening of e-mail archives for confidential information, as well as the ability to place restrictions on content. ePADD screens e-mail for potentially sensitive regular expressions such as credit card numbers and Social Security numbers. ePADD also provides a customizable lexicon tool to aid in searching for and restricting or de-accessioning other types of confidential information.

Archivists can also upload a list of e-mail addresses and restrict access to (or de-accession) any messages associated with those addresses. In addition, ePADD supports the ability to review messages according to various criteria (such as the occurrence of certain correspondents, entities, or lexicon-matching terms) and restrict or de-accession those messages individually or collectively.

Providing Access to E-mail Collections

The same tools that in one context can assist donors, curators, and archivists with reviewing an e-mail archive to enact donor restrictions and manage privacy issues can also support advanced analytic research techniques for researchers.

To mitigate privacy concerns that might make repositories hesitate to make materials available online, ePADD is able to automatically redact from the e-mail archive contextual information including e-mail addresses, message headers, message attachments, and nonentities contained within the messages. Archivists can then generate a redacted copy of the corpus for search, browsing, and visualization on a public discovery website. This means a researcher at home or in the office can still search extracted named entities within a collection, even if the underlying message content is not online. This allows the researcher to determine whether a trip to the reading room to view the full e-mail corpus is warranted. Once a researcher has arrived in the reading room, he or she can access additional functionality within the application. This additional functionality includes the ability to create searchable and exportable annotations for individual messages; to view all image attachments (linked to their accompanying messages) within the application; and to create complex, tiered, thematic searches using the customizable lexicon tool. The lexicon tool enables users to easily perform sentiment analysis and other linguistic analyses across the e-mail corpus and supports visualization of the resulting data over time. Researchers can then add the annotated e-mail messages to a cart that can be requested from the repository depending on the terms of the collection and the policies of the institution. Exported messages can be natively delivered in MBOX file format, a ubiquitous format for e-mail, enabling a researcher to import the messages into a favorite existing e-mail application.

The research support features of ePADD provide a far more robust and interactive way for researchers to discover collection content than using traditional online finding aids. Typical finding aids provide minimal description of e-mail correspondence, such as “Box 5: E-mail correspondence on hard drive.”

What Comes Next?

ePADD has been awarded a three-year IMLS grant to enhance its functionality. The grant begins in November.
2015. The grant focuses on meeting two primary goals. The first goal is to make critical functional improvements to the Appraisal, Processing, Discovery, and Delivery modules. The second goal is to ensure broad and sustained community engagement.

To achieve Phase 2 goals, the ePADD development team and partners will

1. Promote ePADD’s integration within an ecosystem of processes and workflows supporting e-mail ingest and preservation;
2. Build cross-collection and cross-institution discovery capabilities to improve accessibility of e-mail archives to all users in the United States;
3. Facilitate national access to processed e-mail archives approved for public release;
4. Advance ePADD’s support for restricting and releasing of materials;
5. Optimize ePADD for archives of up to 750,000 messages; and
6. Build out new features to augment functionality and performance, incorporating planned additional stakeholder interviews and user testing.

The ePADD development team also aims to ensure broad adoption of ePADD through additional user interface enhancements as well as by building partnerships and offering extensive opportunities for training, discussion, and other engagement by the community.

ePADD is now freely available and customizable for those institutions eager to explore how it can assist with appraising, processing, and providing access to e-mail collections. If you are interested in scheduling a webinar or demonstration to learn more about ePADD, or have any other questions, please contact the project team at epadd_project@stanford.edu. You can also follow us on Twitter at @e_padd.

Notes