A Demonstration of ePADD: Computational Analysis Software Facilitating Screening, Browsing, and Access for Historically and Culturally Valuable Email Collections

Extended Abstract

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

ABSTRACT

ePADD is free and open-source computational analysis software facilitating screening, browsing, and access for historically and culturally significant email collections. The software incorporates techniques from computer science and computational linguistics, including natural language processing, named entity recognition, and other statistical machine learning-associated processes. This demonstration will highlight how these processes enable ePADD to support the appraisal, processing, discovery, and delivery of email archives held by archival repositories and other memory institutions, filling an important role in the preservation of these materials.

CCS 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 PHASE 2

ePADD Phase 2 began on November 1, 2015 and will end on October 31, 2018. Funded through an US Institute of Museum and Library Services (IMLS) National Leadership Grant for Libraries, Stanford University Libraries, with partners University of Illinois Urbana-Champaign, Harvard University, University of California, Irvine, and Metropolitan New York Library Council, are advancing the formation of a national digital platform by further developing ePADD, free and open-source computational analysis software that allows individuals and institutions to appraise, process, and provide access to email of potential historical or cultural value. During this grant period, Stanford University Libraries and grant partners will continue to improve the program's scalability, usability, and feature set [1].

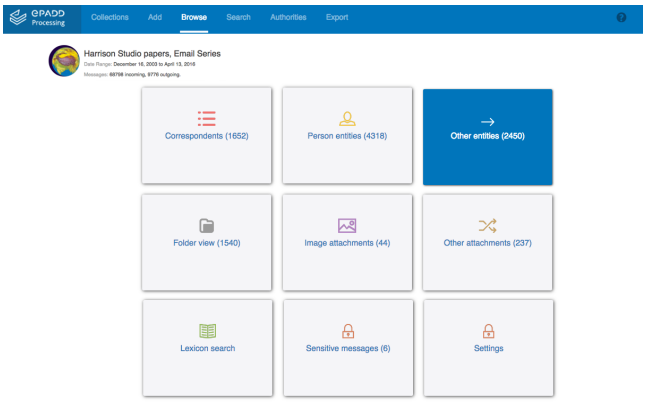

Figure 1: Browse options in the Harrison Studio papers - Email Series, Stanford University, ePADD Processing module, 2017 (Version 3.0).

2 DEMONSTRATED MODULES

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 NER. 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 authorities (see Fig. 1).

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 reading room workstation.

3 DEMONSTRATED FUNCTIONALITIES

3.1 Named Entity Resolution

ePADD uses a custom fine-grained named entity recognizer/classifier that recognizes categories of entities bootstrapped from DBpedia. These include persons, organizations, locations, government entities, political parties, companies, universities, diseases, and awards. ePADD learns from these categories and is also able to recognize likely entities it has not come across before.

3.2 Name Resolution / Correspondent Browsing

ePADD resolves names and email addresses associated with a single correspondent, improving browsing and visualization. All decisions can be manually overridden using a dedicated interface. Mailing lists can similarly be tagged and optionally consolidated using this functionality. Resolved correspondent names can be browsed and graphed alphabetically or by volume of messages exchanged with the email account holder.

3.4 Lexicon Search

ePADD includes tiered thematic keyword searches geared towards broad analysis of a variety of email collections, including lexicons to identify categories of sensitive correspondence. These lexicons can be edited and tuned, or the user can create all new lexicons to suit their research goals.

3.4 Advanced Search

ePADD includes an advanced search interface enabling sophisticated search queries. For instance, users can perform a search for messages containing entities from the *disease* entity category, or terms from the *sensitive* lexicon, and further limit this search by mandating that the search should exclude results from a mailing list. In this way a user can create a narrow search for potentially sensitive information to embargo for a specific period of time or to not transfer to a repository.

3.5 Query Generator

ePADD includes a query generator to aid in comparative entity analysis between the archive and any other textual corpus. Matching entities are highlighted and link to message results.

3.6 Bulk Actions and Annotation

ePADD allows the user to apply actions (including marking messages as reviewed, fit for transfer, or fit for embargo) and annotations to sets of messages meeting user-defined criteria, including all messages associated with a given correspondent, all messages from a given date range, all messages containing certain keywords or named entities in the subject or message fields, or some combination of the above.

3.7 Additional Functionalities

ePADD's additional functionality includes features intended to further support screening for sensitive, confidential, or legally protected materials, as well as features intended to support user access to the intellectual content of the messages. These functionalities include: regular expression search, a redacted view of messages, account and folder-level browsing, built-in visualization tools, and image attachment browsing.

ACKNOWLEDGMENTS

ePADD development is managed by Stanford University's Department of Special Collections & University Archives, part of Stanford University Libraries [2]. The ePADD development team is composed of Glynn Edwards, Peter Chan, Josh Schneider, and Sudheendra Hangal. Development work is performed 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 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 March 17, 2017, from Institute of Museum and Library Services: https://www.imls.gov/grants/awarded/lg-70-15-0242-15

[2] ePADD software, 2017. Retrieved March 17, 2017, from Stanford University Libraries: http://library.stanford.edu/projects/epadd