ePADD: Computational Analysis Software Enabling Screening, Browsing, and Access for Email Collections

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

P. Chan
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
pchan3@stanford.edu

G. Edwards
Stanford University
557 Escondido Mall
Stanford, CA 94305
USA
gedwards@stanford.edu

S. Hangal
Ashoka University
Plot No. 2
Rajiv Gandhi Education City
PIN 131028
India
hangal@ashoka.edu.in

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. In this paper, we explain how these processes enable ePADD to support the appraisal, processing, discovery, and delivery of email 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 [2]. During this grant period, Stanford University Libraries and grant partners will continue to improve the program's scalability, usability, and feature set [1, 8].

2 RESEARCH VALUE OF EMAIL

Email offers singular insight into and evidence of a person's self-expression, as well as records of transactions, collaborations, and networks [7, 9, 10]. Email communications of prominent individuals, including politicians, writers, scientists, and scholars, reveal their professional and personal actions, decisions, and creative output, as well as their relationships within society and communities [6]. The appeal of email collections therefore extends beyond historians to all manner of researchers, journalists, and the general public seeking to obtain insight into individuals and their transactions.

3 ACCESS CHALLENGES FOR EMAIL

3.1 Screening Email

A major challenge that many institutions face when trying to provide access to born-digital collection materials is the need to ensure that creator and third-party privacy, and copyright, are protected [4]. Email archives can include hundreds of thousands or even millions of messages; the challenge of screening email collections is compounded when considered at this scale. As manual review is prohibitively time-consuming, institutions require a mechanism to help automate the process of screening for sensitive, confidential, and legally protected information.

3.2 Providing Access to Email's Intellectual Content

Traditional email browsers typically provide only limited search capabilities, and do not permit browsing of correspondents or named persons or organizations within the email archives. They also do not allow browsing of image attachments or support other types of visualization. These limitations make a traditional browser an imperfect tool for staff of archival repositories and other memory institutions wishing to review or describe email archives, and for researchers who wish to access and study them.

3.3 Supporting Discovery of Email Collections

Current tools to support archival description and discovery for email are limited in their ability to convey the intellectual contents of the archive to a researcher. Email in archival collections has traditionally been promoted and made discoverable online through scant description in library catalog records and archival finding aids, that provide little detail to assist researchers in learning the identity of the principal correspondents, or the named entities discussed.

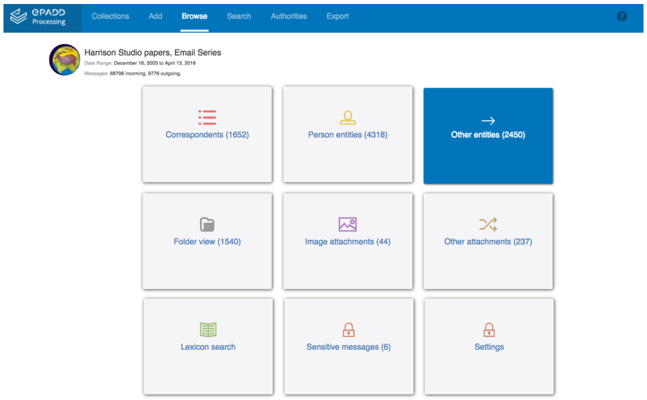
4 ePADD SYSTEM ARCHITECTURE

4.1 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, including mail stored in MBOX format or transferred by IMAP connection. 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.

4.2 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 (see Fig. 1).



Harrison Studio papers, Email Series

Date Range: December 16, 2003 to April 13, 2016

Messages: 68798 incoming, 9776 outgoing.

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

4.3 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 [5]. Only metadata from the processed email archive is published online.

4.4 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.

5 FUNCTIONALITIES

5.1 Named Entity Resolution / Fine-Grained Entity Type Browsing

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.

5.2 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.

5.3 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.

5.4 Advanced Search

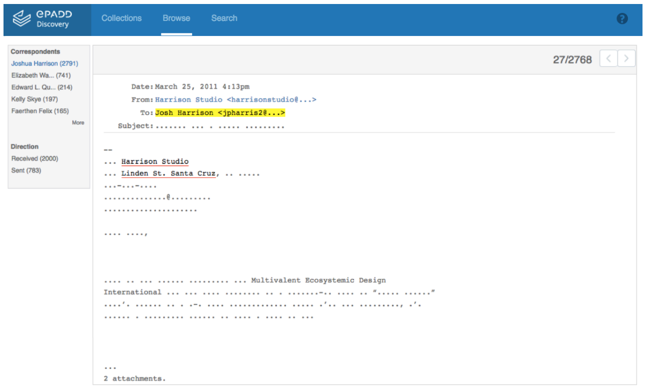
ePADD includes an advanced search interface enabling sophisticated search queries. For instance, users can perform a search for messages containing entities from the *disease* entity category, or terms from the *sensitive* lexicon, and further limit this search by mandating that the search should exclude 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.

5.5 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.

5.6 Redacted View of Messages

ePADD provides an optional Discovery module, intended to provide improved discovery of the archive online via a public web server. This module redacts all content other than message dates, correspondents (local-part of email address), and named entities, in order to protect creator and third-person privacy and copyright (see Fig. 2).



```
      Date: March 25, 2011 4:13pm
      From: Harrison Studio <harrisonstudio@...>
        To: Josh Harrison <jpharris2@...>
   Subject: ....... ... . ..... .........

--
... Harrison Studio
... Linden St. Santa Cruz, .. .....
...-...-....
..............@.........
.....................

.... ....,



.... .. ... ...... ......... ... Multivalent Ecosystemic Design
International ... ... .... ........ .. . .......-.. .... .. ".... ......"
....'. ...... .. . .-. .... ............ ..... .'.. ... .........,  .'.
...... . ......... ...... .. .... . .... .. ...



...
2 attachments.
```

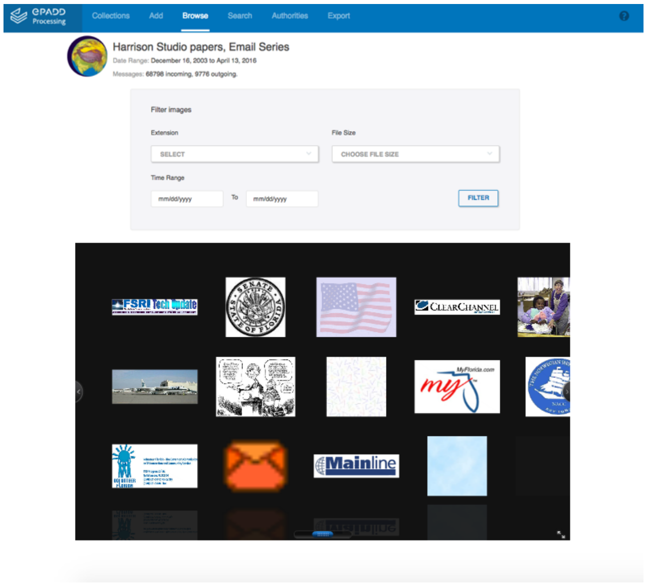
Figure 2: Redacted message view in the Harrison Studio papers - Email Series, Stanford University, ePADD Discovery module, 2017 (Version 3.0).

5.7 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.

5.8 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, account and folder-level browsing, built-in visualization tools, and image attachment browsing (see Fig. 3).



Harrison Studio papers, Email Series
Date Range: December 16, 2003 to April 13, 2016
Messages: 68798 incoming, 9776 outgoing.

Figure 3: Image attachment browsing in the Harrison Studio papers - Email Series, Stanford University, ePADD Processing module, 2017 (Version 3.0).

6 FUTURE DEVELOPMENT

Future development during this grant period includes advancing ePADD's support for restriction and derestriction of materials; continuing to optimize ePADD performance at scale; and development of cross-collection search and browsing. We also plan to develop support for exporting message headers as GraphML for network visualization, and correspondents and fine-grained named entities as linked open data. We will also continue to promote ePADD's ability to support diverse community workflows and institutional requirements.

Long term, we hope to explore development of cross-institution discovery capabilities and the creation of a web service to federate search and browsing across all content that has been processed through ePADD worldwide, in order to streamline the discovery of this content for research.

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. 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 21, 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 21, 2017, from Github: https://github.com/ePADD/epadd

[3] ePADD software, 2017. Retrieved June 21, 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 21, 2017, from Stanford University http://mobisocial.stanford.edu/papers/chi2015.pdf

[5] Harrison Studio papers - Email Series, Stanford University, ePADD Discovery module, 2017. Retrieved June 21, 2017, from Stanford University: http://epadd.stanford.edu

[6] Lukesh, S., 1999. E-mail and potential loss to future archives and scholarship, or, the dog that didn't bark. *First Monday*, 4(9). Retrieved June 21, 2017, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/692/602

[7] Pennock, M., 2006. Curating e-mails: A life-cycle approach to the management and preservation of e-mail messages. *DCC Digital Curation Manual*, S. Ross & M. Day (Eds.), Retrieved June 21, 2017, from http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails

[8] Schneider, J., et al., 2017. ePADD: Computational Analysis Software Facilitating Screening, Browsing, and Access for Historically and Culturally Valuable Email Collections. *D-Lib Magazine* (23:5/6). Retrieved June 20, 2017, from https://doi.org/10.1045/may2017-schneider.

[9] 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

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