Historical Research Using Email Archives

Sudheendra Hangal
Ashoka University
NCR, India
hangal@ashoka.edu.in

Peter Chan
Stanford University Libraries
Stanford, CA 94305 USA
pchan3@stanford.edu

Vihari Piratla
Amuse Labs
Dharwad, India
viharipiratla@gmail.com

Glynn Edwards
Stanford University Libraries
Stanford, CA 94305 USA
gedwards@stanford.edu

Chaiyasit Manovit
Ixora Technology
Mountain View, CA 94040
USA
sit@ixoratech.com

Monica S. Lam
Stanford Computer Science
Stanford, CA 94305 USA
lam@cs.stanford.edu

CHI'15 Extended Abstracts, April 18 - 23, 2015, Seoul, Republic of Korea

ACM 978-1-4503-3146-3/15/04 $15.00
http://dx.doi.org/10.1145/2702613.2702976

Abstract

Archives of letters and documents belonging to individuals provide valuable insights into history. In the digital age, such history is being captured in personal digital archives, especially in the form of email. Archival organizations have recognized the importance of email archives and often collect email when they acquire the papers of eminent donors; however they find it difficult to screen, process and provide access to email for research, due to its sheer volume. We describe the considerations we encountered with the email archives of two prominent individuals in the special collections of Stanford University Libraries. We have designed novel approaches to the challenges of (1) Reconciliation with authority records, (2) Making "finding aids" of the archive available to the general public, without revealing confidential information, and (3) Browsing an email archive when one may not know what exactly to look for.

Our solutions have been implemented in a publicly available and open source system called ePADD. As a result, we enable donors and archival organizations to appraise, process and screen large-scale email archives, thereby unlocking the historical value embedded in them.

Author Keywords
Libraries; Email archives; History.

ACM Classification Keywords
H5.2 [Information interfaces and presentation: User Interfaces.]: - Graphical user interfaces.

General Terms
User Interfaces, Human Factors

INTRODUCTION
As historians know well, letters and documents belonging to individuals serve as invaluable tools of record and provide important insight into the past [7]. Since communication of historical importance is being created and captured in digital archives, archival organizations make it a point to capture "born digital" materials when acquiring the records of eminent individuals. Of these materials, email is perhaps the most significant, due to its long and widespread usage[1], its use for personal communications, and the culture of saving email archives for the record [12].

In this paper, we discuss the issues in making use of these personal email archives in the special collections departments of libraries and archival organizations. We describe and implement specific solutions for the challenges that archivists are likely to face, and illustrate the effectiveness of these ideas with the email archives of 2 individuals – **the poet Robert Creeley** and **the computer scientist Richard Fikes** – that are in our library's special collections. These solutions are built in an open source system called ePADD, which stands for **email: Process, Appraise, Discover, and Deliver.**

Today, email archives are being collected and preserved, but are rarely processed, let alone delivered to researchers and end-users. This is due to concerns about privacy and copyright as well as the difficulty of processing large, multi-decade archives with thousands of messages. While paper records are typically processed manually by archivists, such a process is cumbersome for archives with tens of thousands of email messages. Therefore, **the potential of email archives remains under-tapped** and they are often listed as a single series or sub-series in a "Finding Aid" in special collections, with no further information about their contents. This fact makes it very hard for researchers to make practical use of the archievcs.

Email archives have also become valuable sources of public information. For example, journalists routinely acquire email archives via Freedom of Information requests or from other sources. The email archives of Sarah Palin and U.S. Supreme Court justice Elena Kagan are prominent examples. The Archivist of the United States, David S. Ferriero, reports that emails have been collected from every U.S. administration since the 1980s, and that the archives in the George W. Bush presidential library include about 210 million email messages [2].

Related work
Vannevar Bush presaged the age of personal archives that could be consulted mechanically with "exceeding speed and flexibility" with his vision of the memex [1]. Several projects in the archives community have

[1] Over 2 billion individuals collectively own about 3.3 billion email accounts, according to the Radicati group [6]

recognized the importance of email archives for historical research and are actively working on defining best processes to deal with them [5, 8, 12].

There has been prior work on systems for mining and visualization of email and other text corpora ([11], [4], [3]); of these, only MUSE is publicly available, and ePADD is built on top of it, thus inheriting many of its basic features of visualization, search, summarization and browsing. Projects like Overview and DocumentCloud are popular in the journalists' community for processing textual corpora; however, they are not focused specifically on email archives. Newspapers have attempted to use crowdsourcing to identify interesting information in email archives, for example with the Sarah Palin emails [9, 10].

Archival process

ePADD features four main steps in the process of acquisition and use of email archives. The first phase of **appraisal** involves the donor of the archive (or an archivist) performing an initial screening to decide what is to be preserved. Embargoes and annotations may be be placed on specific items, and sensitive information redacted at this stage. The second phase, **processing**, is for an archivist to examine the archive, clean data, perform another scan for sensitive material, assign authority records, create finding aids if possible, manage any embargoes, etc. This step is time-consuming, taking months, and is often left undone in the absence of tools like ePADD.

The output of the processing phase is the email collection (barring embargoed or redacted material) in a format suitable for discovery and delivery.

The next phase is **discovery**, which allows potential researchers (such as historians, book writers and students) to gain a sense of the content in the archive, e.g., whether certain people or subjects are mentioned, before investing the time and expense of making a trip to the archive's reading room. This step is typically aided with the help of finding aids created in the processing phase. The traditional finding aids for letters in a fully processed collection will typically list the names of correspondents along with the date range of the correspondence; ePADD follows a similar strategy of populating the finding aids with the names of correspondents and entities.

The final phase addressed by ePADD is **delivery**, where the full contents of the archive are available to a researcher, typically in a controlled environment such as a library reading room.

Email vs. traditional correspondence

To illustrate the differences between processing email archives versus traditional paper archives, consider the 7,000 letters in the paper component of the Robert Creeley archive. In the finding aids for this archive, the correspondence listing takes 122 pages out of a total of 251 pages, indicating the importance of letters. Note that this listing had to be painstakingly and manually generated by an archivist going through Creeley's letters. In contrast, Creeley's email corpus consists of 163,689 pieces of email, spanning about 13 years. The messages are loosely organized with relatively little folder structure, and with many duplicates; after de-duplication, the number of messages drops to 49,644. Similarly, in the Richard Fikes collection, there are about 108,000 messages (84,416 unique), spanning a period of about 15 years. The scale of these archives makes it extremely difficult for an

archivist to process each message manually, or for a researcher to examine them individually.

However, email archives have several benefits over physical letters: they can be digitally searched and form a detailed and consistent record over a long period of time that provides a wonderful window into the thinking of the donor [8]. Another advantage of email is that copies of messages often exist with both the sender and the receiver, unlike paper letters where it is more difficult to get a copy of sent letters. This allows chains of conversation to be reconstructed easily. Emails also capture group conversations between multiple people or on mailing lists, and frequently include supplemental images and documents in the form of attachments. The language used in email is quite different from letters – it is frequently informal, colloquial, and uses its own abbreviations and emoticons.

There are also differences due to physical media: physical letters and documents carry useful attributes and signs of authenticity, such as signs of wear or tampering, corrections and margin notes. However, the ease of duplication of digital media is a benefit in some ways because it is easier to preserve the archive over a long period of time.

Appraisal functions

The appraisal phase is meant to be performed by donors, perhaps in consultation with curators from archival organizations. In this phase, email is first loaded into ePADD from mbox format files, or IMAP or POP servers. For other email file formats like Outlook, we recommend commercial tools like Emailchemy or MailStore Home, which can ingest email in a variety of formats and convert to mbox.

A common problem is that personal archives are frequently acquired over multiple rounds of accession spanning many years. This leads to duplication and changes in folder structure. We found this problem in the Creeley archive as well. To tackle this issue, ePADD detects and ignores duplicate messages. We also see cases where some metadata like message recipients or date stamps are missing due to format changes, discrepancies between tools or data corruption. ePADD attempts to deal with these problems as gracefully as it can, by providing reasonable defaults if the data is missing or obviously incorrect. It is common for a single correspondent to have multiple email addresses and name spellings over a period of time; therefore ePADD builds up an address book by merging entities with the same name or email address. This address book can be manually edited by an archivist.

Donors can screen messages for sensitive information that may need to be redacted. Based on our experience with these archives, we have developed a lexicon of terms that are likely to reflect sensitive material, such as student grades, financial or health records, social security numbers, etc. Using this lexicon, the donor can quickly find messages that may be sensitive and can flag them as “Do not transfer” or “Transfer with restrictions”. For messages to be transferred with restrictions, the donor can place an annotation on the message specifying the restriction. Typical restrictions may specify “embargo for 20 years”, or “restrict till death of person X”.

Processing functions

In the processing functions, archivists may perform another thorough screen to ensure that sensitive information is removed. They can also create finding aids for the archive and export separate versions of the archive for use in the discovery or delivery phases.

To help in creating finding aids, we extract entities from each email message in three broad categories, Persons, Organizations and Locations, using the OpenNLP toolkit. Users can browse entities of each type, as well as plot time-based graphs of the most frequently occurring ones.

Authority Records

A major feature of ePADD's processing phase is the association of authority records with the email archive. Authority records are unambiguous identifiers for well-known entities that are manually assigned, for example, in the Library of Congress authority files and similar databases. These authority records make it possible for ePADD to emit records in Linked Data or CSV format, allowing automated lookups of topics and people across vast collections.

Identifying correspondents

The first kind of authority record that ePADD tries to resolve is correspondent names. It is often most important for researchers to determine *who* participates in the correspondence. For this resolution, we use the personal names subset of OCLC's FAST (Faceted Application of Subject Terminology) RDF database. This database currently consists of the personal names of 779,094 well-known people, along with name variants for each. The entries are derived from the Library of Congress Subject Headings, and are linked to other databases like Wikipedia, DBpedia and VIAF when possible.

Given a correspondent name, we first look it up in the FAST database. We found that our name matching needs to be robust to the re-ordering of first and last name, presence or absence of middle initial, etc. It is possible for multiple FAST records to match a single correspondent. This happens routinely in the Creeley archive; for example, one of the correspondents is named "Charles Bernstein", and there are two people with this name in the FAST corpus (and on Wikipedia). One of them is a poet and another is a music composer. Which "Charles Bernstein" might the correspondent in this archive refer to? An archivist could manually dig into the histories of each, and try and scan related messages in an attempt to resolve the conflicted reference. However, this is a tedious job for hundreds of entities with ambiguous authority records, and may need domain expertise, which the archivist may not have.

To aid in this process, we designed an entity resolution mechanism in ePADD that takes the entity's overall context into account, and offers a ranked list of suggestions. All the archivist needs to do is to confirm the authority assignment or select one from a set of candidates, a much easier task.

To provide this list of ranked suggestions, ePADD builds a context for every correspondent that includes all the entities that co-occurred with it in an email message (Single word personal entities like, say, Bob and Jane are not included in the context as they themselves are ambiguous). For each candidate FAST record, ePADD also fetches and reads the corresponding Wikipedia page (obtained from its DBpedia entry, if present). Suggestions are ranked

Figure 1: Each correspondent is resolved to a ranked list of possible matches in the OCLC FAST database.

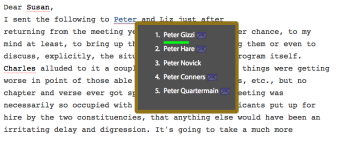

Figure 2: Hovering on a single word name in a message brings up a ranked list of suggested entity completions.

based on the **number of common terms between the entity's context and its Wikipedia page**. To avoid noise in context matching, we found it necessary to exclude the external links and references sections of Wikipedia pages, as they often have incidental and broad terms such as ”CNN”. Our simple matching scheme is frequently able to resolve entities correctly; for example the poet “Charles Bernstein” is correctly offered as the first choice in the example above (Fig. 1).

To evaluate the efficacy of this approach, we compared ePADD's ranked suggestion list to what we manually determined to be the correct authority record on a random selection of 50 correspondents. In the Creeley archive ePADD's highest ranked suggestion was correct 85% of the time.

Disambiguating entities in messages

The above section described how correspondents in the archive are resolved with respect to well-known authority records, disambiguated using Wikipedia text. A different problem is that when individuals communicate, they frequently use common and ambiguous names for entities, which are nevertheless clear to them from their shared context. How might we use context to guess whether a reference in a message to *Mary* refers to cousin Mary or colleague Mary? ePADD offers some help when the user hovers on single-word person names (Fig. 2). It builds a list of candidates for expanding this single-word name by considering all address book entries or person entities in the archive that contain the word. It then ranks these candidates based on the entities and correspondents in the current message, which may contain some clues to resolving the name correctly.

For example, the name *Bush* in a message that also contains the entity *White House* is more likely to refer to the entity name *George W. Bush* than other names that contain the word *Bush* (assuming that *George W. Bush* and *White House* frequently occur together elsewhere in the archive). This feature can be very useful to resolve single word names and help a researcher to quickly make sense of unfamiliar names. However, it is noisier than correspondent lookup in established databases, because a single message may not offer adequate disambiguation context. In our experiment with 30 randomly chosen single-word names that we manually verified, we found that 64% of the highest placed candidates were correct.

ePADD uses algorithms similar to those described above to also resolve place and organization names. For these two entity types, it uses Freebase instead of FAST, as we found that Freebase contains more records and links to Wikipedia for these entity types. Due to space restrictions, we omit details in this paper.

Discovery functions

The discovery module allows partial access to archives over the web. ePADD uses a browser-based user interface even when it is running locally on the same computer. This makes it relatively easy to provide much of the same functionality over the Internet. However, there are two main concerns in the discovery mode: confidentiality and scalability.

Confidentiality: Most archives cannot be made public due to the sensitivity of email messages, issues about copyrights for image attachments etc. A key goal of ePADD is to allow partial (and of course, read-only) access to the archive, publicly over the Internet.

We base the discovery functions on a simple

observation: the traditional finding aids in fully processed collections list the names of all correspondents. Hence, when preparing an archive for the discovery mode, we only consider named entities, and index them as if they were the entire content of the message.

This version of the archive allows users to search for names in the archive, to browse messages with timestamps and names of correspondents (exact email addresses are never exposed) and to view and search for named entities in messages. However, the entire body of messages is not shown. In our experience, this is a good compromise – displaying the names lets researchers generate leads into the archive without revealing confidential details. Interested researchers can follow up by visiting the reading room to get full access. The public server never contains the original email corpus in raw or indexed form, thus ensuring that sensitive data is not lost even in the extreme event of a server compromise. In the discovery mode, a message's source folders and email attachments are hidden; only the number of email attachments is displayed.

Scalability: Normally, ePADD runs on a local computer and only a single user accesses one archive at a time from one instance. We enhanced ePADD with the ability for many users to share access to the same archive in discovery mode. This is important to support many concurrent users (a likely scenario especially when there is some breaking news related to the archive), without requiring a linearly scaled up ePADD server.

Delivery functions

The reading room offers full access to the archive that has been processed by an archivist. A researcher in the reading room can view the full contents of messages with all attachments, similar to the archivist mode except that redaction and export of messages is not possible.

One problem we have frequently encountered in long-term email archives is that messages may include attachments in legacy formats such as WordPerfect, Wordstar and Lotus 1-2-3, and most users do not have the program to open them today. To address this problem, we allow attachments to be viewed with an external viewer (such as QuickView Plus, a commercial tool) for legacy file formats that is installed in the reading room environment. While browsing messages or attachments, users can click on an attachment to optionally open it in this legacy file viewer.

Browsing Features

ePADD supports many of the same functions for visualization and search that MUSE does. However, a researcher browsing an archive that they may be unfamiliar with may not even know where to begin searching.

To aid in this task, we let users perform **a bulk search of terms** extracted from some reference text. For example, to look for interesting messages in the Robert Creeley archive, one could go to Mr. Creeley's Wikipedia page, look for the terms with which he is most prominently associated, and search for them in the archive. We essentially automate this process by letting the user paste in arbitrary text in a search box; we then identify all the named entities within this text

and look them all up in the archive. We display the original text, highlighting terms that had some matches in the archive. In this way, the user can quickly see which of these terms hit in the archive. Further, clicking on a highlighted term leads to the set of messages in the archive that contain the term. This is an efficient way of checking whether the archive has any connections to any entity in the text provided by the user.

Limitations and Future Work

Currently, ePADD can smoothly handle personal archives with about 100,000 messages. Our future plans are to improve scalability and to provide cross-collection search so that library patrons can search multiple collections at once.

Conclusions

We have shown how long-term email archives can be processed relatively efficiently, and how they can be made partially available to the general public. Our experience with the Creeley and Fikes corpora and the resulting system should be useful to other people who need to process large-scale email archives. The ePADD system is publicly available at the URL http://epadd.stanford.edu.

We hope tools such as ePADD will make it more common for curators to capture email archives as valuable documents of record. Currently, this process is limited by the cost of acquisition, processing and delivery.

Acknowledgements

We thank everyone who has contributed to the ePADD project. This project is funded by the U.S. National Historical Records and Publications Commission, NSF Expedition grant CCF-0832820 and the Stanford Mobisocial Laboratory.

References

[1] Bush, V. As We May Think. *Atlantic Monthly 176* (1945).

[2] David S. Ferriero. A New Presidential Library, 2012. http://blogs.archives.gov/aotus/?m=201203.

[3] Hangal, S., Lam, M. S., and Heer, J. MUSE: Reviving Memories Using Email Archives. In *Proceedings of UIST-2011*, ACM (2011).

[4] Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., and Lian, X. Interactive, topic-based visual text summarization and analysis. In *Proceedings of CIKM '09* (2009), 543–552.

[5] Mellon Foundation. AIMS Whitepaper. *In preparation* (October 2011).

[6] Radicati Group Inc. Email statistics report, 2010-2014. http://www.radicati.com/?p=5282.

[7] Shaun Usher. *Letters of Note*. http://www.unbound.co.uk/, 2012.

[8] Susan Thomas. Paradigm Academic Advisory Board Report. *John Rylands University Library, Manchester* (Dec. 12, 2005).

[9] The Guardian. The sarah palin emails, 2011. http://www.guardian.co.uk/world/sarah-palin-emails.

[10] The New York Times. The palin e-mails, 2011. http://projects.nytimes.com/palin-emails.

[11] Viégas, F. B., Golder, S., and Donath, J. "Visualizing email content: portraying relationships from conversational histories". In *Proceedings of CHI '06*, ACM (2006).

[12] Wright, M. Why the British Library archived 40,000 emails from poet Wendy Cope. *Wired* (May 10, 2011).