Processing Email Archives in Special Collections

Sudheendra Hangal
hangal@cs.stanford.edu Computer Science Dept., Stanford University, CA, USA
Peter Chan
pchan3@stanford.edu Dept. of Special Collections, Stanford University Libraries, CA, USA
Monica S. Lam
lam@cs.stanford.edu Computer Science Dept., Stanford University, CA, USA
Jeffrey Heer
jheer@cs.stanford.edu Computer Science Dept., Stanford University, CA, USA

1. Introduction

Libraries and scholarly institutions often acquire the archives of well-known individuals whose work and life has significant historical or research value. Email collections are now an important part of these archives– the Academic Advisory Board Members in the Paradigm project ranked them the most valuable from among images, speeches, press releases, personal websites and weblogs, campaign materials, engagement diaries, presentations, etc [3]. The detailed record embedded in email provides access to the donor's thoughts and actions at a level that has rarely been available in the past and enables researchers to probe questions like: What was the process the donor used to come up with a particular breakthrough? What were they reading at the time and how may it have influenced them? [4] Further, these archives are being accumulated not just by famous people; with about 2 billion users, email reaches just about every section of wired societies. Indeed, the British Library has collected sample email messages from ordinary Britons as a way of capturing a sense of life in the 21st century [5].

In this paper, we describe a new technique for processing email archives in special collections using MUSE (Memories USing Email), an email browsing and visualization system developed at Stanford University. The technical details of Muse are covered in a separate paper [1]. While Muse was initially designed for individuals to browse their own long-term email archives, we have added features that help archivists in processing email archives of others as well. To illustrate with a concrete example, we report our experiences with using Muse to process the email archives of noted American poet Robert Creeley, whose archives are hosted at Stanford University Libraries. The video at http://mobisocial.stanford.edu/muse/creeley.mp4 demonstrates MUSE running on the Creeley archives and supplements the descriptions below. MUSE is publicly available at the URL: http://mobisocial.stanford.edu/muse.

2. Challenges in Processing Email

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 relative difficulty of processing large-scale archives with conventional email tools. While paper records are scanned and processed manually by archivists, such a process is cumbersome for archives with tens of thousands of email messages. Hence the potential of email archives for research remains under-tapped and they are often listed as a single series or sub-series in a "Finding Aid" in special collections, making it hard for researchers to make practical use of them. We elaborate on these challenges below.

Stakeholders

There are several stakeholders in the process of acquisition and use of email archives: the donor, the curator, the archivist who processes the collection, and the researcher who uses it. Each of these stakeholders has different requirements and expertise.

Donors are sometimes hesitant to turn over their email archives to curators as they may contain deeply personal information such as family or financial records, confidential letters of recommendation, health matters, etc. Donors are often busy people and may not have the time to perform a detailed assessment of their archives. Further, a donor may sometimes not be the creator, but say, a family member. Curators develop library collections and maintain relationships with donors.

Archivists are generally well versed in tools and archival processes, but may not be subject matter experts. While archivists want to provide broad access to the archives and encourage exploratory use, they also have to be cautious due to embargoes established by the donor, privacy considerations and copyrights restrictions.

Researchers may be familiar with the subject, but may not be experts with tools. Typically, they would like to gain a sense of the content in the email correspondence through the process of exploratory browsing. They may want to know if certain people or subjects are mentioned in the archives even before making a visit to the collection or raising funding for a project.

Data gathering and cleaning

It is common for digital archives to be acquired at different times, over several rounds of accession, and to be scattered across a variety of digital media including floppy disks, Zip drives, CDs, DVDs and hard drives. Email archives are no exception and we find that, over time, donors change computers, accounts, email clients etc, and store email in different formats (such as Eudora, Outlook and mbox). We have found tools like Emailchemy (http://www.emailchemy.com) useful to convert email in disparate formats to the mbox format that Muse can read. Individuals' email foldering practices tend to be inconsistent over time, and messages are frequently duplicated in various folders. MUSE takes care of this problem by detecting and eliminating duplicates. MUSE also organizes messages by automatically inferred (but manually editable) groupings of people in the archives, making the folder

structure less critical. Further, email addresses and name spellings for the same person tend to change over time; therefore, MUSE performs entity resolution to try and merge records for the same individual.

In the Creeley archives, there are about 80,000 emails; after removing duplicates, MUSE is left with 40,038 messages. Of these, 14,770 are outgoing messages and 25,268 are incoming messages. Creeley corresponded with about 4,000 people in these archives.

MUSE displays graphs of email communication activity, which show that most messages in these archives are from 1996 to 1998, and from 2001 to 2005 (when Creeley passed away), with a sudden dip at the beginning of 2002. This tells us that the archives are missing material from the years 1999 and 2000, and possibly for some period in early 2002. Such signals are useful to the archivist to know that some information may have been missed at some step in the archival process.

Capture and Authenticity

A major benefit of digital records is that they are easy to capture and store compared to paper records. Thus it is possible for a donor to retain access to his or her records for many years or decades, and for archivists to capture the archives of many more individuals and store them in a reasonable amount of space at reasonable cost.

Another benefit of email messages is that, while it is difficult to get access to letters sent by a donor, email copies frequently exist with both the sender and the receiver, leading to a more detailed record. Physical correspondence also has problems with completeness of information. For example, in the Republic of Letters project, many letters were not dated. In contrast, email messages have an automatic timestamp. While both paper and digital formats can decay physically over time, it is easier to preserve a large volume of digital data.

The techniques to determine the authenticity of a paper document are well established. Paper or vellum can be appraised against a familiar set of physical characteristics, such as ink, handwriting, letterhead, paper quality and signs of tampering. However, there are new problems with electronic records. The Paradigm workbook cited above points out that the capture process itself can alter the perceived creation date, and that author metadata is often inaccurate or misleading. Further it notes that establishing intellectual property rights is a key concern for the digital curator who will need to determine who authored a photograph or article, whether they are still alive, whether they still hold copyright and how long that copyright will last.

3. Cues Provided by MUSE

We ran MUSE on the Creeley archives, and found the following cues useful in gaining a quick overview of the archives.

1. Calendar view of terms. MUSE displays a calendar view of the 30 most important terms per month based on statistical ranking, with a novel time-based TF-IDF metric. The terms scored are named entities extracted from the messages using the Stanford NLP toolkit (http://nlp.stanford.edu). We found this feature useful to give ourselves and potential researchers a high-level sense of the contents of the archives; at the same time, the small number of terms makes it easy for the archivist to manually ensure that they are appropriate for public distribution.

2. Sentiment Analysis. MUSE uses sentiment analysis techniques to identify messages that may reflect certain categories of sentiments including emotions (such as love, grief, anger, etc), family events, vacations, congratulatory messages, etc. We have developed these categories and word lists for personal archives instead of relying on more general lexicons like LIWC [2]. See Fig. 1 for a graph of these sentiments over time in the Creeley archives. The MUSE lexicon can be tuned by the user by adding or deleting words to a category, or entire categories themselves. One use we found of this feature was to add a category to identify potentially sensitive messages involving health, finances, recommendation letters, etc.

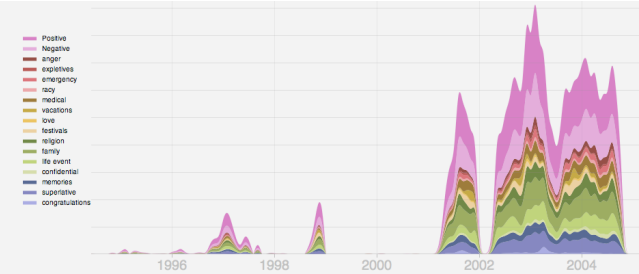


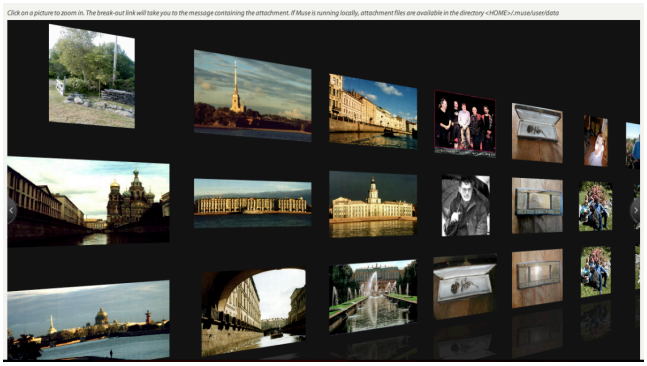
Fig. 1. Sentiments over time in the Robert Creeley email archives.

3. Attachment wall. To facilitate rapid scanning of picture attachments in email, Muse displays them on a 2.5D zoomable and draggable photo wall (Fig. 2). In the Creeley archives, there are 6,282 picture attachments, of which 4,769 are unique. The archives include many interesting pictures, for example, those of Creeley and his family, his home, trips he took, various forms of artwork, and scanned announcements of events. In general, pictures and documents have associated copyrights, so the archivist cannot publish these images and attachments publicly.

Browsing features

When the user is interested in following a cue, Muse launches a message view with all the messages related to that cue. These views can be fairly large and consist of hundreds of messages. To make it easy to rapidly skim a collection of messages, MUSE provides a faceted browsing interface, where the facets are sentiments, groups, people, original folders, email direction (incoming vs. outgoing), and month or year. It also provides a jog dial interface that lets users rapidly flip through messages without the need for keypresses and mouse clicks. The jog dial is very popular with users of MUSE.

Fig 2. Image attachments in the Robert Creeley email archives.



Multiple views

While we initially thought that Muse would be used primarily by archivists, we realized that it can be useful to donors and researchers as well. To support these stakeholders, two distinct views of the archives are needed. The first is a full-access view for donors and archivists to use when processing the archives. The same interface can also be made available to researchers in a reading room environment. The second, more limited interface can be made public and can provide enough detail for potential researchers to get an overall sense of the archives' contents. It can include a calendar view of important terms, and perhaps the overall patterns of communication with different groups and sentiment. However, the actual message contents are omitted. Since Muse stores message headers, bodies, and attachments separately for each folder in its own cache, we found it easy to support both views for the Creeley corpus; in the public view, we simply hide the message bodies and attachments.

Message selection and export

We envision that Muse can be used by donors themselves to screen their archives before turning them over to the library, with the help of features like automatic grouping of email and sentiment analysis. We have added a feature in Muse to allow users to tag messages and export all messages with a particular tag (thereby including only the selected messages), or without a particular tag (to redact certain messages). These features can also be used by an archivist to screen the archive for sensitive material.

4. Connecting the Archives to Web Browsing

While scanning the Creeley archives through Muse, we realized that it would be useful to look in it for terms for which Robert Creeley is most famous. This is a difference from the original purpose of Muse; the archivist or researcher is not expected to be intimately familiar with the life of the donor. For example, according to his Wikipedia page, Creeley is known as a Black Mountain poet; searching for this term in his archives returns 259 messages. We therefore implemented a browser plug-in that searches for named entities on the page being browsed and highlights those that are also present in the archives. Clicking on the highlighted text lets the user explore email messages that include the term. This lets a researcher bring the archives' lens into his normal browsing.

We hypothesize that researchers can find this feature useful to browse their own research. The browsing lens will automatically find terms in the archives that are relevant to the researcher's interest and highlight them.

5. Conclusion

Our overall experience of processing email archives using Muse was quite positive. Muse can help archivists by letting them spot missing or unclean data, performing quick scans of the contents for material that needs to be restricted, and make parts of the archives publicly available for researchers' use. Researchers benefit by gaining an overall sense of the material in the archives; when they need to drill down into the actual contents, an interactive browsing and navigation interface aids them explore the archives efficiently, and a browser plug-in lets them bring a lens from the archives into their normal browsing.

Using Muse, archivists can hope to process email archives quickly and make valuable information available for researchers. Further, we believe that making Muse even simpler to use will enable ordinary individuals to browse their own long-term email archives, or those of people close to them such as family members. This will allow the study of personal archives on a scale that has not been possible until now.

Acknowledgements

We thank Glynn Edwards, the Andew W. Mellon Foundation, NSF POMI 2020 Expedition Grant 0832820 and the Stanford Mobisocial lab for supporting this work.

References

1. S. Hangal, M. S. Lam, and J. Heer. Muse: Reviving memories using email archives. In Proceedings of UIST 2011. ACM, 2011.
2. LIWC Inc. Linguistic Inquiry and Word Count. http://www.liwc.net.
3. Susan Thomas. Paradigm Academic Advisory Board Report. John Rylands University Library, Manchester, Dec. 12, 2005.
4. M. Wright. Why the British Library archived 40,000 emails from poet Wendy Cope. Wired, May 10, 2011.
5. S. Zjawinski. British Library Puts Public's Emails on The Shelves. Wired, May 29, 2007.