CoSA-NHPRC Symposium
Government Email in an Age of Risk:
Preventing Information Loss
September 15, 2017

Case Study 6:
Redaction Issues in Email Access
Texas State Library and Archives Commission

Brian Thomas, Electronic Records Specialist, Texas State Library and Archives Commission

Background

As the official repository of state government records, TSLAC receives records from state agencies (and, to a lesser extent, retired legislators who choose TSLAC as their repository of record) in all of its forms; and agency records retention schedules rarely differentiate between how paper and electronic formats are to be handled. Following good records management practices, email records are most often received (as of 2017) by TSLAC in the same way that paper counterparts would be received, as loose electronic files amongst other electronic files of a larger set of records. As a result, the majority of emails received by TSLAC are individual message files (in the Outlook .msg format) rather than entire inboxes (such as an Outlook .pst format).

TSLAC manages its electronic records through the Texas Digital Archive (TDA). The TDA was established during Fiscal Year 2015 in anticipation of the transfer of the Perry administration records in January 2015. The repository platform used by the TDA, Preservica, allows for automated preservation workflows such as migration on-demand as well as public access to unrestricted records. Since the TDA was established, TSLAC has received records from Governor Perry, several legislators, the lieutenant governor and several state agencies. In addition to new transfers specifically for the TDA, electronic records have been received on an ad-hoc basis as part of transfers of analog materials. As of the end of June 2017, of 2,246,964 files totaling 37.72 TB there are over 67,000 Outlook email messages, 29 Outlook email inbox archives files, and 11 iCal files in the TDA. This is excluding materials that have not yet been ingested into the repository and new incoming materials. Projections for these variables are not available.

There are several factors that play into the issue of email redaction at TSLAC. The primary factor is the Public Information Act (PIA), which has a complex set of rules about what is and is not releasable as part of public records. The rules regarding what is public information are highly contextual and require a nuanced understanding of the law to be able to effectively determine what is restricted. While automated tools do exist to identify some types of restricted information, such as social security numbers, none of the tools can apply the PIA provisions properly.[1] Another factor in this issue is that we require evidence, at the point of redaction in a file, that a redaction of content occurred.[2] To date we have not found any tools that can leave this kind of trace while maintaining the email in an email file format. The best tool for redaction we have found is Adobe Acrobat Pro, which only works with PDF files. Given that redaction must occur with a PDF file, a final factor is transformation of emails to PDF for redaction. Our preservation repository can store the email files in many formats and extract metadata to aid in tracking down records, but is unable to perform any transformations to PDF files. Thus, conversion from email to PDF must be done manually by archivists downloading records from the repository and using an Outlook plug-in on the individual files.

Work methods and Scope

Unfortunately, due to the volume of materials and the varying nature of restriction rules, it is not possible to spend dedicated time during or after initial processing to systematically go through the email files in our collections for restricted information and redact that content. There is insufficient staff time for this level of processing and it is not possible to delegate the task to volunteers. Thus, the scope of email redaction is limited to redaction by an assigned archivist on-demand for a public information act request.

Work on email redaction takes the following steps:

a. Locate responsive (as in applicable to the request) email records using various means. This includes full-text searching the preservation repository for migrated records, full-text searching of secondary local copies of these same records, and (in the case of the Governor's records) searching the correspondence tracking systems available for logged messages. Located message are downloaded and copied to a local folder for processing.
b. Transforming responsive messages to PDF for redaction. This step involves importing the email(s) into an archivist's Outlook inbox, so they can use the Adobe Acrobat plug-in to transform the email to a portfolio PDF. The portfolio format permits retention of email attachments. These portfolio PDFs are created either en-mass for a set of emails (produces one file) or one per message.
c. Redacting PDFs. Using the tools available in Acrobat, archivist redact restricted information from the email text.

[1] For example, names of juvenile offenders in the Department of Juvenile Justice records are restricted, but the staff working with them are not. Since the names of juvenile offenders vary from record to record there is no way to reliably determine what names to redact without reading the document. Another example is email addresses of private citizens, which are restricted in some (but not all) cases and may share the same domain name as a business (i.e. both use Gmail).

[2] Imagine a black line on a page like in the days of redacting paper records.

d. Redacting attachments. Attachments that require redaction are also transformed to PDF, then Acrobat redaction tools are used to remove restricted information. Attachments that do not require redaction but are in an easy to change form, such as from office productivity software, are also transformed to PDF. All transformed attachments are attached to the appropriate email in the portfolio PDF and the original attachments are removed. It is worth noting that after this step, any clickable links in the email text to transformed attachments will no longer work.

Issues and challenges
We have found many challenges with our handling of email redaction, the following are the most noteworthy:

a. Conversion of emails to PDF using Adobe Acrobat, and most other tools, does not preserve some of the header information, such as IP addresses and transmission data. Is provision of that type of information, which will almost always be restricted, necessary or even appropriate? Or is it the content of the transmission that is important?
b. Without additional tools to rely on, searching has been restricted to text searches of the loose email files. Given that it is significantly easier to search an entire email inbox using the Outlook program than to search loose email files, is it appropriate to aggregate loose files into an artificially created email archive for archivists to use while processing Public Information Act requests?
c. The directory structure a responsive email was gathered from is not recreated when the records are released to the requestor. When providing access to redacted emails, is it necessary to recreate the directory structure the email came from to let end-users know the context?
d. Using a basic directory search does not search email attachments for potentially responsive items, but searching through Outlook does. Is this sufficient reason to aggregate loose items into email archive files?
e. Multiple emails and folders of emails can be converted to portfolio PDFs as one combined file, is this an appropriate way to represent a larger structure?
f. How is email defined? If a PIA request were to ask for email to a public official, is the responsiveness to the request strictly based on the format type (i.e. as defined by the Network Working Group standards RFC2045/RFC822) derived from the standards-based email transfer protocol? Since more and more agencies are using web form submission over traditional email (including the U.S. legislature), does that electronic communication fall within the scope of email?
    i. This classification question is critical to determining whether things like the Governor’s correspondence tracking system files falls within the email category and email-centric metadata needs to be assigned to them. That would be an increase by probably hundreds of thousands of electronic correspondence items.

ii. This question is also germane to archivists dealing with correspondence management systems that automatically convert received email into database data, like Lockheed Martin's IQ product.

Results (problems and lessons learned)

a. When searching for responsive email records, review your work carefully. With records both in the preservation system and awaiting processing to be ingested into the preservation system, it may not be as simple as searching a single location for materials for redaction.
b. It is possible to emulate email inboxes during conversion using the PDF portfolio approach. This is only scalable to the folder/directory level as email is too voluminous to do at the inbox level.
c. Attachments in a PDF are not always readily visible in a converted email, although there are indicators of where an attachment should be. It is not ok to assume that just because an email text is unrestricted that its attachments are unrestricted. Or that those attachments do not need to be migrated to PDF format.
d. Email as a standard has very specific rules that allow for transmission between email systems regardless of platform or source format. Email storage formats do not share this level of inter-operability. Open source tools for working with emails as developed in academic environments rely on open source formats like .EML and .MBOX. These are supported by some email client software applications and web-interfaces. However, the overwhelming majority of businesses and state agencies use Microsoft Outlook as an email client, with its proprietary .msg and .pst file formats. This begs a utility question for many of these tools.

Thoughts for further work in this area and outstanding questions

a. In preservation terms, how important is it to keep header information when redacted versions of the email will not include that data?

b. Based on the email projects we reviewed, the industry presumption seems to be that all emails will be received by an archives as a inbox from an individual/office, even though good records management necessitates filing email individually based on where it goes in a retention schedule. Why is that? How can we improve the existing tools that work with email for identifying potentially sensitive information to include other relevant format types, such as the random individual email file?

c. Although itself proprietary, TSLAC's preservation system relies largely on open-source tools to transform files to new formats. There are no open-source tools that can migrate MSG or EML files to PDF, nor are there tools to transform PDF attachments to PDFs themselves. As an ubiquitous format that is not easily modified by public users, PDF is a common dissemination format. Is this an area where further tool development is needed in the open-source community?

d. Traditional email comes in a variety of file formats but complies with very specific transmission and header rules. Web forms used by companies or agencies in lieu of providing direct email addresses contains some, but not all, of that information. The underlying concept, standardized text-based electronic correspondence between parties, is the same. Should we be talking more about electronic correspondence as a whole, or reframing the discussion in those terms?

Summary

Over the last several years the Texas State Library and Archives Commission has received electronic correspondence ranging from loose email and email archive files to web-form generated correspondence. None of these formats are easily processed for identifying restricted information and visualization using current (known) popular open-source tools. The answer the redaction of proprietary email formats for us has been transformation to an entirely different type of proprietary format, portfolio PDF. This resolves the redaction issues in the immediate future, but sheds light on some problems with current viewpoints on handling email. Current processing tools are focused on the idea that email transfers should be an entire inbox, how do we balance this with the reality that retention schedules can call for emails as part of a larger scheme?