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

Case Study 3:
Tests and Examples of Internal/Commercial Tools
Illinois State Archives Email Preservation and Access

Brent West, Assistant Director, Records and Information Management Services, University of Illinois System

Background

The Illinois State Archives (ISA), established in 1921, serves as the depository of public records of Illinois state and local governmental agencies which possess permanent administrative, legal, or historical research value. Its collections do not include manuscript, newspaper, or other nonofficial sources. These records are available to the public, officials, and scholars at the Norton Building in Springfield, Illinois and at seven regional depositories located on state university campuses throughout Illinois. The Archives provides access through a series of printed and electronic guides, and by in-person, mail, telephone, fax, and Internet database reference services.

The Archives' responsibilities also include oversight for the management of state and local government records retention schedules. Illinois has a well-established legal framework for the management of government records of all types, and has successfully overseen the efficient management, disposition, and preservation of government records for many years. Until recently, this process has been entirely paper-based. However, with the increased use of electronic records by government agencies at all levels, the ISA has made great strides over the last several years to update and adapt their procedures and processes to accommodate records in this new environment.

In recent years, the Illinois State Archives has established a strong partnership with the University of Illinois for the purpose of exploring options for jointly operating a digital preservation repository. Starting in 2014, the Archives and the University began a four-year project using Preservica Cloud Edition to assess that software's viability as a long-term

repository solution. Results have indicated the Preservica option meets expected requirements for digital preservation at a cost that should be sustainable.

Project methods and scope

Background and scope

Although the project began in earnest in January 2017, preliminary conversations began in early 2015 to identify and acquire historically-significant electronic correspondence from senior Illinois public officials that are deemed permanently valuable by established retention schedules and Illinois law. By late 2015, the current administration granted permission to the Illinois State Archives to transfer email from 62 high-value accounts associated with three former gubernatorial administrations going back to 1999. In 2016, the ISA and University partnered with researchers with the National Institute of Standards and Technology's (NIST) Text Retrieval Conference (TREC)[1] to evaluate several machine learning tools against a subset of this email which had been coded for several topics of archival interest. These and related efforts to evaluate e-discovery tools for archival purposes led to our current project.

The project focuses on identifying, securing, processing, and providing access to high-value email messages in a reasonable time-frame. Projects such as the Library of Virginia's Kaine Email Project[2] have shown that manual processing can be done on email collections, but the resources needed to accomplish it are time intensive and unsustainable from both a budgetary and manpower perspective. The goal of this project is to explore the use of predictive coding tools in addressing the challenges associated with processing large digital collections, applying them to email collections from state agencies in a sustainable way, and ultimately make the collections appropriately available for researchers and the general public.

Methods

We began by applying the National Archives and Records Administration's (NARA) Capstone email approach, treating accounts of senior State officials as having email worthy of retention based on their position. For our project, we are focusing on the email messages of persons working in the Office of the Governor for the State of Illinois. Individuals from the following positions were identified: Governor, Deputy Governor, General Counsel, Chief of Staff, Deputy Chief of Staff, Director, or Senate Liaison. The email administrators gathered the email of 32 individuals from the original list of 62; nearly half had already been lost. After removing duplicate messages, this 320 GB dataset was reduced to approximately 166 GB or 2 million messages.

Our project was divided into discrete phases based on the kinds of effort we envisioned. The phases we mention in our project proposal are designed to focus on the minimum tool/script/system functionality needed to "process" a set of email messages so that we can make it publicly available through the Illinois State Archives reading room. The phases of the project were originally set up as:

[1] http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf, Accessed 7/17/2017

[2] http://www.virginiamemory.com/collections/kaine/, Accessed 9/15/2015

1. Acquiring static batch of email (including preservation)
2. De-duplication tools assessment (including text extraction)
3. Auto-categorization tools assessment
4. Restrictions/Redactions tools assessment (including removal of non-archival items)
5. Enhancement tools assessment
6. Batch Email Processing
7. Search and Access tools evaluation

We are working with graduate students from the School of Information Science with computer science skills and an interest in text mining to review both open source and commercial tools that may provide viable workflows for future archivists. Once we got our work underway, it became apparent that our work plan needed some flexibility, which is described in the next section.

Issues and challenges
Are the stages of the project lined up in the right/best order?

As we have gotten into the project, we are learning that it is not practical to have clearly separated phases based on data processing steps such as de-duplication and restriction/redaction. Some of the commercial tools package various tools into one complete suite or system. Therefore, we are now focusing on a more thorough review of specific systems or tools regardless of what functionality they may have that aligns with our original project phases.

- *Difficulty negotiating contracts with vendors through state agencies*

Because the University of Illinois and the State of Illinois are public bodies and subject to procurement rules aimed at protecting public funds, the process of negotiating contracts with vendors is very cumbersome and time consuming. We need direct access to specific products for assessment purposes and the grant funds are available to support a modest level of commercial product exploration. However, the timeframe to finalize contracts requires advanced preparation and attention to detail. In addition, some vendors have been difficult to contact or slow to reply. Others have been unable to accommodate a novel use case such as this at an affordable price.

- *Determining the best method to provide access*

We are not currently focusing on how we will ultimately provide access to the processed content, but this may be more challenging than originally anticipated. Will there be expectations from users that they should be able to search across all the email at once for particular concepts or key words? Should we pre-populate folders with content we believe relevant to popular searches that will be anticipated or should we allow full capabilities of whatever tools/software we end up using available to end-users? What sort of workflow do we imagine will allow us to easily add more email to what we already have made available at the

end of this project? How will restricted messages be managed for limited access, on-demand redaction, and/or future availability?

Problems solved, successes, failures, and lessons learned
The problems we have encountered so far have more to do with administrative procedures and operational start-up matters. Most of these problems have been resolved through ongoing communication. One of our project objectives is to provide recommendations for a sustainable workflow going forward. Therefore, we need to work with the state IT personnel to identify and document a simple process of securing batches of email annually. It may take several rounds of securing email batches prior to assuming we have a good process.

A definite success we can identify is the partnership we have developed between the Illinois State Archives and the University of Illinois. We have also developed good working relationships with experts familiar with some of the tools we are exploring.

We have no particular failures or lessons learned to report yet.

Questions remaining, ramifications or further work or research in work/project
Once we have assessed various tools, we will need to address issues related to identifying confidential information. Quick wins should be addressing the obvious sensitive matters such as personally identifying information, matters associated with personal health issues, and/or attorney-client privilege communications. Less obvious issues of confidentiality may not be easily identified upfront. We are hopeful the tools will allow us to make short work of some of the processing steps so that more time may be spent reviewing content for more complex confidentiality issues such as student disciplinary incidents or other private matters that may need restricted access for some period of time.

Other questions we will need to address have to do with how to incorporate other digital content the Archives may be acquiring with this email content. What will be the best method of providing access to researchers? How much should the Illinois State Archives be expected to do to make contextual connections for researchers versus how much should that effort be left to the researchers themselves? As technology continues to evolve, email will likely become overshadowed by social media venues. When should the Illinois State Archives be expected to capture various instances of postings to social media? What kind of agreements should be developed between various archival organizations, whether they are federal, state, academic, or private, to leverage skills and aggregation efforts? Academic libraries are reducing redundancy in their physical collections, relying on designated entities to hold and preserve just a few copies of books once held by many. How might such a model translate (or not) to the archival community?

Summary
In summary, we are excited to continue to press forward with this project to see how some of the tools that have already been developed by the e-discovery community might be repurposed to help archivists review of large quantities of digital content, particularly email. Our work is

focused on these sorts of tools with an understanding that they may help archivists save time in reviewing content that they are considering for long-term preservation and they may also help researchers review and access the content once it has been preserved. We still have a long way to go, but we think the partnership we have developed and the approach we are using demonstrates practicality and can lead to a sustainable model for preserving digital content for the future.