Electronic Currents

Assistant Editor: Joanne Kaczmarek, University of Illinois. Contact Joanne at jkaczmar@illinois.edu if you would like to guest author a column or have a good idea to share.

TOMES: A State Archives Story about E-mail

By Camille Tyndall Watson, Digital Services Section Head, State Archives of North Carolina

TOMES: Transforming Online Mail Using Embedded Semantics (TOMES) is a three-year grant that has partnered the State Archives of North Carolina, the Utah State Archives, and the Kansas State Historical Society to begin investigating and solving the problems presented by archiving e-mail in a state government setting. The project, which began in October 2015 and will be completed in September 2018, has two parts: developing a methodology for implementing the Capstone approach in archiving state government e-mail accounts and developing an open source natural language processing (NLP) tool with dictionaries specific to state government to assist with the processing of large government e-mail accounts.

First Step: Capstone

The project began by setting up a collaboration among the records analysts in the Government Records Section (GRS) of the State Archives of North Carolina with the purpose of identifying state agency personnel positions that could be considered Capstone positions. Capstone is an approach to managing government records developed by the National Archives and Records Administration (NARA) in 2013.[1] Capstone allows government agencies to streamline how they manage e-mail by categorizing and scheduling e-mail at the account level based on the function and/or position of the e-mail account owner.[2] A series of forms developed by NARA was adapted by project team members and state archives records analysts to be distributed to agency chief records officers for the purpose of collecting information about personnel positions. The first form provided information on positions that would be identified as Capstone based on their position in the organizational chart. The second and third forms provided information on positions whose functions would make them likely to contain archival e-mail, regardless of their place in the organizational structure. Information collected included the position title, the names and e-mail addresses of the people who had held the position for the past five years, and, most important, the HR position number. The HR position number is a unique identifier used to automate reports identifying people leaving Capstone positions to facilitate the transfer of those inactive Capstone e-mail accounts. The Office of State Human Resources and Department of Information Technology (DIT) were instrumental in creating these reports. The e-mail accounts of positions identified as having the Capstone designation will be placed on hold until the accounts can be transferred, allowing DIT to follow the state's five-year retention period for non-Capstone e-mail, as per Governor McCrory's Executive Order 12,[3] while retaining permanently valuable e-mail accounts for transfer to the archives.

Next Step: Building Tools

The TOMES project is also building a natural language-processing tool that will help state government archives address the challenges associated with processing and providing access to e-mail accounts identified as public records. This work includes developing a workflow in Docker that will convert e-mail accounts to EMLs before being run through an adapted instance of DarcMail to convert the EML to EAXS, the XML e-mail schema developed by the State Archives of North Carolina as part of the 2009 grant-funded project, EMCAP. The EAXS preservation file is then run through the TOMES tool to be tagged for Personally Identifiable Information (PII) and other confidential information, as well as named entities. The TOMES tool will also generate a METS file built for describing e-mail accounts. The final Archival Information Package (AIP) for each e-mail account, as currently conceived, will contain

- The original e-mail file (PST, MBOX, EML)
- Untagged EAXS file
- Tagged EAXS file
- METS metadata file
- Statistics on the e-mail account

This is admittedly a large AIP, particularly when dealing with large government e-mail accounts; but as no way currently exists to validate the account between file transformations, a large AIP seems appropriate. This challenge presents an area of research ripe for development in the future.

The TOMES tool is built so that the tables feeding into the tool for tagging and building the METS profile can be easily edited for each state's needs and public records laws. The tool and documentation are still in development, but the code is publicly available for testing and comments on the State Archives of North Carolina's GitHub page.[4]

Once the e-mail account has been through the TOMES tool, it is considered "processed." However, to provide access to the public, the tags will need to be reviewed, so a workflow for iterative processing based on patron requests is in development. This iterative processing approach will allow public access to e-mail accounts over time, while being realistic about the demands on the archivists processing the e-mail. The tagging will allow archivists to identify where PII and confidential information may be found in an account. An elastic search is being developed to allow archivists to search an account based on NLP tagging. After individual messages are reviewed by the archivist, they can be marked as available, restricted, or redacted, and made publicly available accordingly.

Final Step: Successes

By the end of the project, a free, adaptable, and easy-to-use tool for the MPLP processing of large e-mail accounts for mediated access will be available. However, more work is still to be done, and project partners are hoping to do that work with support from a future grant cycle. Most state agencies provided responses to questions about Capstone e-mail accounts. However, the appraisal criteria need to be further defined. Once complete, the refined appraisal criteria will help solidify and clarify those positions designated as Capstone. The final Capstone list will require active, ongoing conversations between agencies and records analysts, as well as ongoing communication as organizational structures change over time.

The TOMES tool as it will be at the end of the current grant will begin to give archivists a great deal of useful information when processing e-mail accounts, but there is more work to be done. The project partners would like to improve the accuracy of the tagging and begin experimenting with machine learning technology to improve the tagging of e-mails and provide public access to them more quickly. Additionally, while the current TOMES tool can identify materials that need to be redacted, building out the redaction functionality of the tool so that redacting can be done more easily is a future goal. The project team would also like to build out an access module that can load up an e-mail account and only provide patrons access to e-mail messages within those accounts that have been redacted and approved for public access.

Future Chapters

Since the TOMES grant began, members of the team have been engaged in conversations around the country regarding e-mail archiving. Conversations have included those sponsored by the NHPRC Email Symposium in Washington, DC, and the Mellon Foundation Email Task Force. Through these conversations, it has become clear that e-mail is a growing concern throughout the archives community and that the tools to come out of TOMES could be of use beyond the state government context. The project team encourages future grants to build on the products of the TOMES project, engaging the academic and corporate archives communities to see how the tools could be adapted for those environments. Additionally, the wider digital preservation community may find the TOMES project of interest as an iterative processing approach is developed to make e-mail available to the public. Expectations are that the approach might change over time as the AIP in the OAIS model will be changing, potentially over the course of several years, as the e-mail account is processed, making it no longer a static, fixed digital object.

The TOMES project has been an exciting and illuminating process to work on through the myriad questions that e-mail archiving presents to the archival profession. TOMES project partners look forward to exploring the new opportunities presented by the project and sharing lessons learned. TOMES documentation can be found on the Council of State Archivists portal for electronic records resources database[5] and reports, resources, and partner contact information can be found on the TOMES project web page.[6]

Notes

1. NARA Records Managers, Email Management, NARA Email Management Requirements, https://www.archives.gov/records-mgmt/email-mgmt.
2. NARA Records Managers, NARA Bulletin 2013—02 "Guidance on a New Approach to Managing Email Records," https://www.archives.gov/records-mgmt/bulletins/2013/2013-02.html.
3. State Library of North Carolina, North Carolina Digital Collections, "McCrory, Pat Executive Order No. 12 Amending the State Email Retention and Archiving Policy," http://digital.ncdcr.gov/cdm/compoundobject/collection/p16062coll5/id/20571/rec/12.
4. State Archives of North Carolina GitHub page, https://github.com/StateArchivesOfNorthCarolina.

(Continued on page 16)