North Carolina Department of Natural and Cultural Resources

ROY COOPER, Governor **SUSI H. HAMILTON, Secretary**

TO: Nancy Melley, National Historical Publications and Records Commission

FROM: Camille Tyndall Watson, Project Director, Digital Services

SUBJECT: NAR15-RG-50005-15—Transforming Online Mail with Embedded Semantics Collaborative Project

DATE: October 1, 2017

The Transforming Online Mail with Embedded Semantics Collaborative Project is a multi-state, multi-part collaborative initiative to identify state positions by function that are responsible for generating or receiving archival email, develop methodologies to move archival email accounts out of their native system, and to develop predictive coding and libraries to assist archivists to process those email accounts and make it available. The project team includes Camille Tyndall Watson, Jeremy Gibson, Jamie Patrick-Burns and Nitin Arora (NC), Elizabeth Perkes (UT), and Megan Rohleder (KS), with staff from the Library of Virginia contributing and offering their insight and experiences. The project also has an advisory committee—Cal Lee, University of North Carolina at Chapel Hill and Chris Prom, University of Indiana at Urbana Champaign.

Process at least ten email accounts designated as containing permanently valuable correspondence based on Capstone roles.

Staff from the State Archives of North Carolina (SANC) have successfully been able to pull down complete email accounts from the Office 365 cloud-based eDiscovery module. We have successfully transferred email accounts from the Perdue administration, and Governor McCrory's email accounts. We have prepared an export of email accounts from the rest of the McCrory administration, and are working with DIT to complete the download; due to the large size of the files, we are planning to load the export directly from DIT servers to an external drive for transfer.

SANC staff have continued to connect with state agencies to update their responses to the Capstone identification forms. We have broken the .csv file created to track responses into separate files based on agency, to track historical data. We also developed a workflow for continuing to update documentation of Capstone position.

Staff from the Department of Information Technology (DIT) and Office of State Controller (OSC) in North Carolina have continued to investigate how to use the Capstone position number information gathered by records analysts to automate the "tagging" of accounts in Microsoft Outlook and BEACON, the state's personnel tool. Staff from SANC staff have met several times with staff from DIT, OSC, and NC Unified Communications to discuss our email archiving requirements and work through the workflows needed to automate the tagging of Capstone accounts with a legal hold, including what the trigger will be for SANC to begin negotiating the transfer with the agency that has custody of the email account. Most recently, SANC staff met with OSC staff separately to decide how the Capstone variable will be coded into Beacon, and what report can be generated from their systems to allow SANC staff to determine when an employee separates from a Capstone position. We hope to have some sample reports by the end of the year, and functionality with Office365 by the end of the grant.

As we have begin to conceptualize a workflow for collecting and processing email, we have come to understand that effectively processing email accounts for completely open access will require one or more pass at the records by an archivist. This has led us to consider a new, theoretical processing and access paradigm: iterative processing. We have started to work with Records Description Unit staff at SANC to discuss processing workflows and how the TOMES tool will work with their existing methods. Currently, the Records Description Unit frequently processes records "on demand" as there is patron need for records, due to the large amount of backlog in government records holdings. On reflection of these workflows, we felt that iterative processing using the TOMES tool was a more realistic concept. NLP and feature extraction can only take us so far in "processing" emails. Due to the overwhelming number of emails processing will need to happen in stages with the NLP tags and other machine derived statistics providing guide posts to records within an email account. The first pass will be to identify emails with potentially sensitive information. These will be flagged by machine, but verified by a processing archivist. Over time as emails are accessed, the machine's tags will be verified or modified, and gradually this will produce a fully processed account. While the grant deliverable discusses the processing of 10 accounts, we understand that it will take an iterative approach to make them fully available, due to existing workflows within archival institutions, in addition to the need for effective removal of PII and other confidential information. The TOMES project will see us implement the first part of the iterative processing approach, which will allow for greater ease in access related to patron requests, even if further processing will be required for full access to an individual account. We hope to have suggestions for further work beyond the grant period.

Produce a cross platform .pst to EAXS XML parser

The cross-platform PST to EAXS converter in the Tomes project is DarcMailCLI. It is now stable and feature complete. We used it to convert multiple .eml and .mbox files successfully into EAXS. Work on DarcMailCLI currently consists of bug fixes and minor enhancements as issues arise during the NLP processing of the extracted messages.

The current stable version is hosted on our Github account:
https://github.com/StateArchivesOfNorthCarolina/DarcMailCLI

Publish an NLP dictionary designed to flag named entities unique to government at the state and local level

The main thrust of development over the past 6 months has been the NLP and “semantic” portion of the tool.

We have resolved our initial question of “which NLP libraries are TOMES going to utilize?” as we have incorporated Stanford's CoreNLP suite into our entity tagging codebase. CoreNLP was chosen over other software we investigated because we found it to have the best combination of documentation, institutional and community support, and applicability to our objectives with TOMES. Specifically, CoreNLP offered the best route for utilizing software already trained to identify entities while offering us the ability to customize the output with our own entity patterns, including those that require the use of regular expressions such as social security numbers and other forms of personally identifiable information. Regarding our own entity patterns, we are building our entity dictionary as well as the software required to read the dictionary from an Excel file and "feed" the customizations to CoreNLP.

At the end of September, the grant team in NC met with the Government Records Section staff from Records Analysis and Records Description to begin collecting name entities and PII patterns that are specific to NC state agencies. Staff from the Records Analysis Unit will be identifying tags that are specific to the agencies they work with, while Records Description Unit staff will be identifying tags based on their needs while processing archival email accounts. We hope to have a completed survey of these terms by the end of December 2017. In October, we will be meeting with the team members in Utah and Kansas to collect additional entities and PII patterns that are applicable in their state governments to make the NLP dictionary more broadly applicable across state governments.

To store the semantic email data, we have devised an XML output format based on the original EAXS schema. It is from these "tagged messages" that statistical information and searchable data could be extracted to assist an archivist in future processing of the message.

We have also made progress in devising methods for the identification of signature lines in emails. Our message tagging work has led us to experience the unique issues that signatures lines pose for NLP-based entity detection in emails. Signature lines are often equated to "noise", redundant data often with little relevance to the email message in question. As we were unable to locate existing tools to extract signatures from emails, we hope that our work can lead to future improvements with semantic tagging of email.

The current tagging-related toolset is hosted on our GitHub account:
https://github.com/StateArchivesOfNorthCarolina/tomes_tool

Publish findings, tools, and training pieces on a TOMES project website, through the CoSA Program for Electronic Records, Training, Tools and Standards (PERTTS) portal, social media platforms e.g. Twitter

Camille Tyndall Watson presented the progress of the TOMES project at the Society of American Archivists (SAA) meeting in Portland, OR in July 2017. The presentation, discussed the project’s history, the process of identifying permanent email accounts, collaboration with DIT and OSHR to automate the tagging of identified “archival” positions, and progress on the processing tool. She also discussed final, expected deliverables for the grant, as well as the benefits and drawbacks of a Capstone approach and challenges we have encountered in developing the tool. She also introduced our concept of iterative processing of email accounts based on access demand.

In September 2017, Camille, Jeremy Gibson, and Sarah Koonts also attended the NHPRC-CoSA sponsored symposium Government Email in an Age of Risk: Preventing Information Loss in Washington DC. At the symposium, Camille presented a white paper entitled "Working with Stakeholders to Create/Influence Policy", which addressed the team's experiences, including challenges, successes, and lessons learned, implementing a Capstone email retention policy in NC. Camille presented on the TOMES projects at Chief Records Officer meetings in April and September to report on the progress of implementing Capstone in NC, and to get CRO feedback and buy-in.

The SANC team has had calls with the Alabama Department of Archives and History to discuss how we have implemented Capstone, and the Task Force on Technical Approaches for Email Archives to discuss the challenges we have found in email preservation, and how the TOMES tool works. NC has also been approached by NAGARA to participate in a webinar about the project in March 2018. Other states have also requested and have been sent our documentation, most frequently documentation on implementing Capstone.

Future Work

The TOMES team will continue to refine and expand the NLP processing of email bodies. The primary focus of this development period will be building out state-specific NLP dictionaries, engaging archivists to design feature sets that will allow them to identify significant records as they process. We have already asked archivists in NC to begin working on terms for the NLP dictionary, and we will have a meeting of all state partners in October to develop the NLP dictionary for use beyond NC state government. We will also investigate methods for archivists to review email accounts that have been tagged by the TOMES tool.

North Carolina has purchased 224 TB of storage to store archival email accounts, and will receive and install the new server in the next 6 months. Storage purchase was delayed due to issues with IT procurement, both within DNCR and the larger Department of Information Technology. However, we placed the order before the end of year two, and expect to have the storage provisioned soon.