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: April 30, 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' process that mail and make it available. The project team includes Camille Tyndall Watson, Jeremy Gibson, and Nitin Arora (NC), Elizabeth Perkes (UT), and Megan Rohleder and Ryan Leimkuehler (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.

Design a methodology for identifying positions in state government whose email represents an asset of significant historical value (similar to NARA's Capstone approach) and "tagging" the email accounts associated with those positions for periodic ingest.

Staff from the State Archives of North Carolina (SANC) continued to connect with state agencies to record their responses and complete forms 1 & 2. We were successful in receiving responses to all three forms from the Department of Transportation (DOT) and Office of State Controller (OSC) in 2016. To keep track of the information we received from the agencies, staff compiled a comma separated value (.csv) file with the position number, position title, name on account, name of account and the agency. To keep the .csv file up to date, records analysts reached out to agency CROs again this year in anticipation of the annual state CRO meeting in April. We hope to receive any updates or changes to the forms, as well as additional information from agencies that did not return all three forms last year. We will also be developing a workflow for continuing to update documentation of Capstone position.

Staff from the Department and 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 met with DIT staff to discuss our email archiving requirements and how their proposed solutions may work, and next steps to achieve our goals. We will also be working with DIT and OSC to develop workflows for the transfer of email accounts as people move in and out of Capstone positions.

Additionally, NC records analyst Kurt Brenneman published a case study on the implementation of the Capstone forms for *Information Management* magazine, a publication created by ARMA International.

Utah is currently reviewing and adapting the forms for use. Kansas plans on evaluating and implementing the forms once their transition to a new email provider is completed at the end of May.

Transfer of email accounts

The NC Digital Archivist, Camille Tyndall Watson, began pulling down email accounts from the North Carolina Office365 cloud using Microsoft's eDiscovery tool and its export command. The exports are currently set to create one .pst file per mailbox, to include any unindexed materials, and to deduplicate the .pst. The exports also include an export summary .csv, an XML manifest, a .csv summary of the export, a log of the export, and a .csv of any unindexed items. With the election of Governor Roy Cooper, we were able to identify and pull down Capstone accounts from the Office of the Governor from Governor McCrory administration, as well as Capstone at-will appointments from the Department of the Natural and Cultural resources that changed with the change of administration.

We continue to export these emails as one .pst file per account, but are attempting to export administration emails as one packaged export containing the .pst files, since this is more efficient process that takes advantage of eDiscovery's search capabilities. However, there have been some problems with downloading the export packages for larger administrations, like Governor McCrory's. We are working with DIT staff to create a workflow that will allow us to get the email accounts in a timely manner. We also continue to address other questions that had come up before the election, such as empty mailboxes.

Since the last report we have identified several suitable methods for removing formatting html embedded in email. The primary reason for removing formatting html from the emails is to eliminate noise prior to using NLP tools to process the emails, as such, this is a key piece of the final tool. Each method has been successfully applied to our test data sets. Currently we are evaluating each for speed and accuracy. Another need for the tool is to leave formatting html such as <a> tags, we are also evaluating each tool to make sure they can accomplish this task accurately.

Utah provided North Carolina with email harvested from Gmail accounts from one local government and several state offices: Town of Apple Valley, State Treasurer, Department of Technology Services, Department of Administrative Services, and the Department of Health. These accounts were transferred in the form of .eml and .mbox files. The email accounts were primarily from executive officers such as department directors and deputy directors.

Hiring a Programmer

In November, we were forced to let our two programmers go, due to scheduling conflicts. Also in November, we hired a new programmer, Nitin Arora, who has been working with us full time. Since

November, we have been working on the TOMES tool in two spaces: 1) A rewrite of DarcMail, 2) NLP tools and text processing.

Carl Shaefer's DarcMail tool provided an initial codebase for the conversion of email accounts in the mbox format to EAXS. For the goals of the TOMES project, however, DarcMail needed some modification. For instance, Utah uses Gmail for their government email. Gmail primarily exports email accounts as an EML file. The TOMES version of DarcMail needed to process these accounts. Another roadblock we found was memory overflows with extremely large accounts. We have forked the original DarcMail adding the needed functionality, and ported the code to Python 3. As of this report, we have successfully converted 6 email accounts into valid EAXS, and 4 accounts with a valid JSON mirror.

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

The other line of development has been the NLP and "semantic" portion of the tool. We have made significant progress in the last 6 months. Our grant funded programmer has been working to isolate semantically significant portions of the email to increase the signal to noise ratio inherent in large email accounts. The main issues that we have addressed are: 1) How do we identify and exclude/include signature lines from NLP consideration; 2) which NLP libraries are we going to utilize; 3) incorporate our own state specific rules into the NLP process to tag state specific passages or entities.

The current toolset is hosted on our Github account: https://github.com/StateArchivesOfNorthCarolina/tomes_tool

We have been developing software to identify Personally Identifiable Information (PII) within emails and conversion tools to transform HTML-based emails into accurate plain text representations.

Devise, test, and implement an email preservation workflow to ultimately transform email to the Email Account XML Schema (EAXS) utilizing the services and hardware of their e-mail system

Jeremy and Nitin have also been working together to devise, test, and implement an email preservation workflow to ultimately transform email to EAXS and package the EAXS files with accompanying attachment files and preservation metadata. In designing the functionality of the tool, many workflow items are built into the tool, such as converting .psts, use of Darcmail to convert and run NLP processes, and loading the AIP into a discovery module, using Docker as a candidate deployment platform.

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 Best Practices Exchange (BPE) in Sacramento, California. 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.

In February, Jeremy Gibson presented on the TOMES project at the nlp4arc conference held at the University of North Carolina Chapel Hill. His presentation addressed the scope of the TOMES project as well as the inherent difficulties with government email in terms of the volume of data and the role NLP can play in aiding email archival practices in the 21st century. He discussed the application of NLP within the TOMES project and the necessity of using NLP tools to assist in the processing of large email corpora.

The SANC team has had calls with the Library of Virginia regarding the processing work they have done on the Governor Tim Kaine emails, as well as professors from the University of Waterloo, who have been partnering with LVA to build a machine learning tool that can process email, testing their program's results against the hand-processed account from LVA's processing effort. University of Waterloo was willing to share their published materials on the tool, and are willing to collaborate with us once our tool is further along.

Additionally, SANC staff have shared their experiences with the grant through the Council of State Archives (CoSA) State Electronic Records Initiative (SERI) program. In January, Camille Tyndall Watson took part in a webinar aimed at IT staff on how IT and archives staff can work together to improve digital preservation; in her presentation, she discussed SANC's ongoing collaboration with DIT on the tagging and transfer of archival emails. We have also begun working to post the Capstone account identification forms onto the PERTTS portal for wider dissemination to other state archives.

The Next Six Months

The TOMES team will continue to refine and expand the NLP processing of email bodies. The primary focus of this development period will be identifying features of email signatures to assist, testing those features, and evaluating their effectiveness. Also during this period, we will continue to engage archivists to design feature sets that will allow them to identify significant records as they process. This will focus primarily on a new processing/access paradigm that we think may be necessary for large email corpora that we call "iterative processing". 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. 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 on access, and gradually this will produce a fully processed collection.

North Carolina is planning with DIT for the purchase of storage for the continued collection of archival email accounts. Utah and Kansas will continue to evaluate the Capstone email identification forms, and will begin testing the tools that have been put in GitHub by the North Carolina team. Kansas plans on evaluating and implementing the Capstone identification forms once their transition to a new email provider is completed at the end of May.