North Carolina Department of Natural and Cultural Resources

PAT McCRORY, Governor **SUSAN KLUTTZ, Secretary**

TO: Nancy Melley, National Historical Publications and Records Commission

FROM: Kelly Eubank, Project Director, Digital Services

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

Date: October 31, 2016

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 Kelly Eubank, Jeremy Gibson, and Camille TyndallWatson (NC), Elizabeth Perkes (UT), 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). We are still contacting other agencies for information regarding the third form. In order 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. Looking forward, SANC discussed how to make this information available. After reading the Records Express blog, SANC investigated the feasibility of hosting the information in Github and making it public. However, staff decided to table the discussion on this until it receives feedback back from OSC regarding the using the position numbers in Office365 and the account.

SANC staff continued to pursue methodologies to automate workflows to “tag” accounts. We talked with groups from Department of Information Technology (DIT) and Office of State Controller (OSC) to describe what we are trying to do and why. In our discussions, both DIT and OSC staff were open to finding a solution. Essentially, in order to move forward, there have to be modifications in the personnel management software. It can be done but the software will have to be modified. Staff at OSC is willing to pursue the modifications pending legal and IT approvals. Staff at OSC took this to the agency legal group to verify that position numbers are public record. The agency chief information officers (CIO) are also reviewing these efforts and providing feedback.

In Utah, because they have requested to harvest the email accounts of directors who are leaving, the Department of Human Resources have started alerting the Utah State Archives when a director is leaving. It is still not as automated as they must rely on people letting them know but it is an improvement over looking for newspaper articles.

Transfer of email accounts

The digital archivist 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. To date, we have successfully downloaded three accounts from identified positions of the Office of the Governor during the Bev Perdue administration (2009-2012). These exports range from 2-9 GB, and we are currently preparing to pull down more accounts.

The export process has also allowed us to identify challenges for the transfer of email accounts, as well as challenges for processing the accounts once they are exported. Some email accounts exist in the eDiscovery system, but the system does not detect any messages. For example, one account appears to have 2 mailboxes connected to it, but no messages. In these cases, the account cannot be exported for further examination. We are working with DIT to understand and resolve this issue.

Additionally, newer email messages exported from the tool that contain HTML code, which does not render after the .pst has been put through our tools. This causes many emails containing HMTL to be almost unreadable. Because there does not appear to be a setting in eDiscovery to avoid including the HTML in the exports, we hope to find an alternative solution to making these emails accessible.

Hiring a Programmer

Beginning in March 2016 the TOMES team began our search for a programmer. We developed a job description with core competencies required. Our budget for a programmer was $56,000 per year for two years. As this salary is well below market rate for a seasoned developer, our strategy was to find Computer Science students or recent graduates from the Universities in the Raleigh/Durham area. Before applying for the grant, we talked about this idea with our IT group to gauge if it would be a successful strategy. DNCR-IT was enthusiastic about it and offered to help us when we were ready. In early April, prior to graduation, we published our job request to NC State and UNC Chapel Hill’s job service boards, and received over 25 applications. Five of those were suitable for the position, with one outstanding candidate from the School of Information and Library Science (SILS) program at UNC Chapel Hill.

Unfortunately, after we began the hiring process, the Office of State and Human Resources (OSHR) modified their policies on hiring external temporary labor for time limited projects. SANC was no longer allowed to execute a personal services contract but rather had to go through the staffing service the state contracts with for temporary employees. Because the temporary employment agency takes on the administrative pieces of hiring and paying an employee, SANC now had to factor in overhead costs to the agency. This put a halt to our attempts to bring aboard the SILS candidate, while we came to an understanding of how to comply with the new guidance from OSHR. By the time we were ready to move forward again, the SILS candidate had moved on to another opportunity. The NC staff contacted other identified candidates to see if they were still interested; however, due to the long time gap, they had also moved on.

Another consequence of the policy change was that we needed to adjust our salary cap down in order to comply with new overhead costs associated with the policy change. This limited the pool of qualified candidates. After completing these modifications, we reposted our job search in late June, and by mid-July had a candidate –a graduate student currently enrolled in the Computer Science Department at North Carolina State University. Due to his obligations as a student, he could not commit to a 40-hour work week. To solve this problem without incurring even more delays, we suggested that perhaps we could hire two programmers for 20 hours per week. Our candidate proposed a qualified colleague of his, who subsequently submitted an application.

The process of working the two programmers through the OSHR system completed in mid-August. We had meetings immediately following. The programmers made their first reports by September 3. They are currently working on the first steps of converting Personal Storage Tables (PST) to MBOX. There are only a couple strategies available for the conversion of PST to MBOX. One is the use of an existing commercial software solution like Aid4Mail or Stellar. The other is the open source library libpst1. The advantage of the libpst solution is that we can build it into the TOMES tool itself rather than requiring users of the tool to purchase a license for a commercial product and splitting the workflow into multiple processes.

Currently, we have tested libpst on various sizes and versions of PSTs with excellent results. We have processed PSTs ranging from 2GB in size to 14GB in size and have found that libpst is able to accurately reproduce the folder structure of a PST file in MBOX format.

The next goal of the project is to convert the MBOX folders into EAXS. The tool we are testing for this piece is a python program (DarcMail) developed by Carl Schaefer at the Smithsonian Institution. DarcMail walks a file structure of MBOXs and produces an EAXS file and three metadata files. DarcMail additionally allows for the splitting of an email account into "chunks" which mitigates some of the potential access problems associated with having a single huge xml file. During the testing process Jeremy Gibson found that DarcMail makes some assumptions about how the final structure of a processed account should look that could prove problematic for an OAIS compliant TDR. First, DarcMail can store attachments in external xml structures using Base64 encoding for the attachment body. However, in the current version of

[1] http://www.five-ten-sg.com/libpst/

DarcMail these files are stored in the MBOX file structure. The TOMES team considers the MBOX structure to be an intermediate step, and as such would require an additional step to extract the files into a finalized AIP structure. To address this, Jeremy Gibson forked DarcMail and added the ability to move those attachments into a directory outside of the MBOX folder structure. Second, the original version of DarcMail did not de-duplicate email attachments at runtime. The forked version of DarcMail allows the archivist to decide whether to de-duplicate attachments at runtime, reducing the final AIP size in our tests by half.

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

Evaluate EMCAP software for suitability

After evaluating the EMCAP software, SANC determined that the large majority of the EMCAP project deliverables are not suitable for TOMES.

The Server Stack

In order to evaluate the suitability of the E-mail Capture and Preservation (EMCAP) software we worked closely with our partners in DIT, additionally Jeremy Gibson was given access to the original codebase written in C#.

EMCAP has very specific requirements in order to function:

OS

Windows Server 2003 Service Pack 2 (32-bit x86)

Software Prerequisites

Microsoft SQL Server or Express 2005 or later

Microsoft SQL Server Management Studio

Microsoft .NET Framework 2.0

Microsoft Core XML Services 6.0

Microsoft IIS (Internet Information Services) 5.1 or later

hMailServer

DNS Registration

EMCAP requires the server to be registered on a local or global DNS server

The first problem we encountered is that Windows Server 2003 is no longer supported by Microsoft and therefore was generally unavailable for new installs through the Department of Information Technology (DIT). However, we were able to effect a build within a virtual environment along with all of the other software prerequisites. Unfortunately, even with this build we could not get the processes to reliably run as they were intended. These were most likely due to the necessity of having clients connect using name resolution, or some other obscure software incompatibility. The tight integration of EMCAP with outdated MS technology, made it very difficult to diagnose the problem.

Conclusion

EMCAP was developed nearly a decade ago as an Enterprise solution using an Enterprise software stack, and hardware stack. As such, even if we were able to build a functioning server, the technical and business case hurdles to overcome for potential users of the TOMES tool would be extreme. For Enterprise users like State governments, who would have the technical knowledge to install and maintain the EMCAP service, installing software out of service life on trusted networks would most likely be a non-starter. For users with more limited resources like memory institutions or local governments, the highly integrated stack of requirements would most likely prove too burdensome to even attempt.

The EMCAP XML generator

Part of the EMCAP project was an email account schema (EAXS). This schema was intended to model the structure of a typical RFC 2822 (Internet Message Format). For the EMCAP project, XML conforming to the Schema was generated using a series of C# classes. We thought that we could port these classes over to Java or Python as modules to be used in the TOMES tool. While we were evaluating this possibility, it came to our attention that Carl Schaefer at the Smithsonian Institution had already produced a functional tool (DarcMail) to convert files in the mbox format to xml which can be validated using EAXS. As this is an already functioning piece of software, in one of our target languages we decided to abandon attempts to port the C# classes.

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

The Tomes website is now established-- https://www.ncdcr.gov/tomes. Several months back, the Department website was hacked and SANC lost the ability to add or edit websites. DNCR, the parent agency plans to move divisions from their current website to the Drupal template used by the state and we are waiting to hear more about that timeline. As a result, the url could change. We also discussed a twitter account specific to the project. However, after discussion, we decided to use the existing Twitter accounts used by the Division and use the hashtag #TOMES. The social media team will work in coordination to ensure that content is posted on all platforms used by the division—Facebook and Twitter primarily.

At the Annual Meeting of the Society of American Archivists (SAA), Camille Tyndall Watson presented the progress of the TOMES project as part of a panel called “One Year into the NHPRC’s State Government Electronic Records Grant Program.” The panel, which also included staff from Wisconsin and Missouri, discussed the achievements and challenges of each state involved in the 2015 round of NHPRC State Government Electronic Records program. Camille discussed the project’s history, the process of identifying permanent email accounts, and collaboration with DIT and OSHR to automate the tagging of identified “archival” positions. She also discussed the goals for the grant, and the expected deliverables.

From the SAA presentation, the TOMES group connected with a group from Harvard who has their own email archiving tool. That tool is primarily focused on ingest of email accounts. The TOMES group engaged in a conference call with the Harvard team to discuss how our tools could be a part of a toolkit for any institution to use. Harvard will be publishing a paper from the Archiving Email Symposium Workshop. The twitter hashtag for this is “#ArchEmail”. This workshop was co-hosted by the Library of Congress and the National Archives and Records Administration focused on workflows, toolsets, and policies for accessing and preserving email archives and brought together representatives from federal, academic, state, and technological areas to discuss the challenges of trying to preserve and make accessible “archived” email. The group anticipates publishing papers from this conference and Patty Fouchy of Harvard promised to make that available to us. The Harvard tool is not open source as it was primarily built for the environment at Harvard. However, they are interested in making it an open source tool and we will engage again with them once the paper is published and we are further along with our project. All of the participants were very excited about the prospect of putting together a suite of tools.

Going forward

The TOMES team will focus on working on processing concepts beginning with the phenomenal work done by the Library of Virginia on the Governor Tim Kaine email. The programmers are also in place so we will begin developing libraries and testing that against the corpus of email accounts that we have. In addition, we should have an answer about the capability to mark email accounts as archival and have that workflow in place.