Transforming Online Mail with Embedded Semantics (TOMES)

Camille Tyndall Watson
Digital Services Section Head
State Archives of NC
Where We’ve Been...

- EMCAP, 2007-2009
- Executive Orders 150 and 18, 2009
- Mimosa Nearpoint, 2009-2014
Where We Are...

- Email retention reduced to 5 years
- State IT has become centralized
- Email has moved to the cloud with Office 365
- Email archive migration

Image: Google Maps
Transforming Online Mail with Embedded Semantics (TOMES)

• 3 year grant (2015-2018)
• Partnership between State Archives of NC, Utah State Archives, and Kansas State Historical Society
• Advisory group includes Cal Lee (UNC-Chapel Hill), Chris Prom (University of Illinois Urbana Champaign), and staff from the Library of VA
What should we keep?

Capstone Approach

- Permanent - top decision-maker’s e-mail
- Temporary – all other staff, 7 year retention
- Non-record – 1 year retention

Images: British Postal Museum and Archive; National Archives and Records Administration
Building Relationships

- Collaboration with Government Records Section, SANC
- Development of a series of forms to understand organizational structures of state agencies
- Outreach to State Agency CROs and CIOs for education and buy in
### ASSESSMENT OF EMAILS FOR PERMANENT RETENTION

**PART 3**

**Agency:**

Please identify positions that create official records that document agency policies and decisions related to any of the programs or subjects listed below. These programs or subjects are not represented in the accounts of personnel already listed on Parts 1 and 2 but still may need to be retained permanently. Please note: if emails that document the programs or subjects listed below are already being captured by the email account of an agency position listed on the Parts 1 or 2 forms, including instances where the agency executive has been copied on these emails, it is not necessary to list those positions here. For each individual identified below, please list the email accounts of all predecessors in that role since January 2011.

**Possible email subjects with archival value:**

1. Major agency policies
2. Formulation of rules and monitoring standards (e.g., Administrative Code)
3. Events, incidents, and situations that required a prolonged response involving multiple agencies and led or had the potential to lead to large-scale loss of life, severe damage to lands and property, or major disruption of the state's infrastructure
4. Direction and planning of the core program(s) of your agency
5. Cooperation with external state and/or federal agencies
6. Construction and real property transactions
7. Major public events, such as the State Fair, First Flight Centennial, inaugurations, etc.
8. History of the state of North Carolina
9. Advocacy for minorities, such as Indian tribes
10. Certification, commissioning, etc.
11. Management of assets held in public trust for the people of North Carolina – e.g. state parks, historic sites, artifacts, archival materials, etc.
12. Evaluation of rules created by other agencies, where the agency is an established part of the rule review process

<table>
<thead>
<tr>
<th>NAME</th>
<th>POSITION TITLE/ROLE</th>
<th>DEACON POSITION NUMBER</th>
<th>ARCHIVAL SUBJECT MATTER</th>
<th>EMAIL ADDRESS</th>
<th>BEGGINING DATE FOR EMAIL COLLECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Plugging In

- Department of Information Technology
  - “Tagging” accounts by function
  - Facilitating the transfer of email accounts from cloud storage
- Office of State Human Resources
  - Identifying positions by position number
  - Working with DIT to “tag” accounts
Making It Work

- Harvesting email from cloud using Office365 eDiscovery tools
- Created email processing modules written in Python as a toolbox for preserving and processing email accounts
- Developing NLP libraries for tagging
EAXS

- Tagged EAXS file
- Attachments folder
- METS metadata file
- Statistics spreadsheet

TOMES

.pst

Read PST

.mbox

Darc Mail
The Final AIP

• Original inbox file (.pst, .mbox, .eml)
• Untagged EAXS XML file
• Attachment XML file
• Tagged EAXS XML file
• METS file
• Statistics spreadsheet

<SingleBody>
  <Charset>us-ascii</Charset>
  <BodyContent>
    <![CDATA[
      Hi,
      
      Can the TOMES software recognize SSNs?
      For example, would it tag 721-07-4426 as an SSN? Don't worry, I got the number from Wikipedia.
    ]]>
  </BodyContent>
</SingleBody>

Hi,

Can the TOMES software recognize SSNs?
For example, would it tag 721-07-4426 as an SSN? Don't worry, I got the number from Wikipedia.

Thanks,
Fred

ps: I'm enjoying my vacation from Paris.
Challenges

• Capstone and Account identification
  • Development of appraisal criteria
  • Communication with DIT and OSC

• Tool Development
  • HTML in newer emails
  • Signature lines
  • Identifying State Government and State Agency-specific acronyms
  • Developing a useful METS profile
The Final Product

- Development of State Government specific NLP libraries for use in the processing of email accounts containing public records
- An MPLP approach to the arrangement and description of email
- The ability to identify materials that should be reviewed for PII before release to public
- Mediated access using iterative processing
Stay in touch!

• GitHub: https://github.com/StateArchivesOfNorthCarolina
• Website: http://www.ncdcr.gov/tomes
Questions?

Camille Tyndall Watson
Digital Services Section Head
State Archives of NC
Camille.TyndallWatson@ncdcr.gov
(919) 807 - 7359