Transforming Online Mail with Embedded Semantics (TOMES)

Camille Tyndall Watson
Digital Services Section Head
State Archives of NC

Kelly Eubank
Head of Content and Information Delivery
State Library of NC
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
Capstone Approach

Permanent - top decision-maker's e-mail

Temporary – all other staff, 7 year retention

Non-record – 1 year retention
Collecting Data

ASSESSMENT OF EMAILS

Agency: ________________

1. The head of the agency, or equivalent, and all high-level staff
   of the agency.

   For each individual identified:

   a. Role and position in the agency.

   b. Responsibility for receiving and responding to emails.

   c. Description of the issues or topics handled through email.

   d. Frequency of email communication.

   e. Examples of specific email interactions.

   f. Any policies or procedures related to email management.

   g. Contact information for any additional personnel.

   h. Any other relevant information.

   i. Any recommendations for improving email management.

   j. Any comments or observations.

   k. Any other relevant information.

   l. Any other relevant information.

   m. Any other relevant information.

   n. Any other relevant information.

   o. Any other relevant information.

   p. Any other relevant information.

   q. Any other relevant information.

   r. Any other relevant information.

   s. Any other relevant information.

   t. Any other relevant information.

   u. Any other relevant information.

   v. Any other relevant information.

   w. Any other relevant information.

   x. Any other relevant information.

   y. Any other relevant information.

   z. Any other relevant information.

   AA. Any other relevant information.

   BB. Any other relevant information.

   CC. Any other relevant information.

   DD. Any other relevant information.

   EE. Any other relevant information.

   FF. Any other relevant information.

   GG. Any other relevant information.

   HH. Any other relevant information.

   II. Any other relevant information.

   JJ. Any other relevant information.

   KK. Any other relevant information.

   LL. Any other relevant information.

   MM. Any other relevant information.

   NN. Any other relevant information.

   OO. Any other relevant information.

   PP. Any other relevant information.

   QQ. Any other relevant information.

   RR. Any other relevant information.

   SS. Any other relevant information.

   TT. Any other relevant information.

  UU. Any other relevant information.

  VV. Any other relevant information.

   WW. Any other relevant information.

   XX. Any other relevant information.

   YY. Any other relevant information.

   ZZ. Any other relevant information.

   AA. Any other relevant information.
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
The diagram illustrates the process of reading PST files and converting them to EAXS files, with the following steps:

1. PST files are read by the Read PST tool.
2. The output of the Read PST tool is an EAXS file.
3. The EAXS file can be converted to a Mbox file.
4. The Tagged EAXS file can be further processed by Darc Mail.

The EAXS file includes:
- Tagged EAXS file
- Attachments folder
- METS metadata file
- Statistics spreadsheet
The Final AIP

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

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