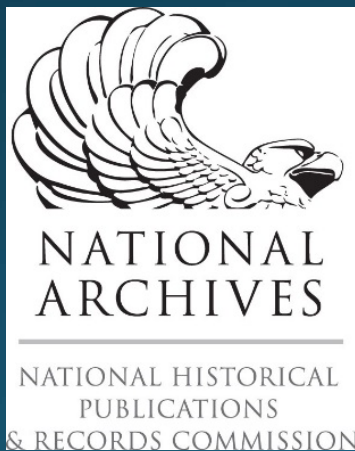


Transforming Online Mail with Embedded Semantics (TOMES)

Camille Tyndall Watson
Digital Services Section Head
State Archives of NC



State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES

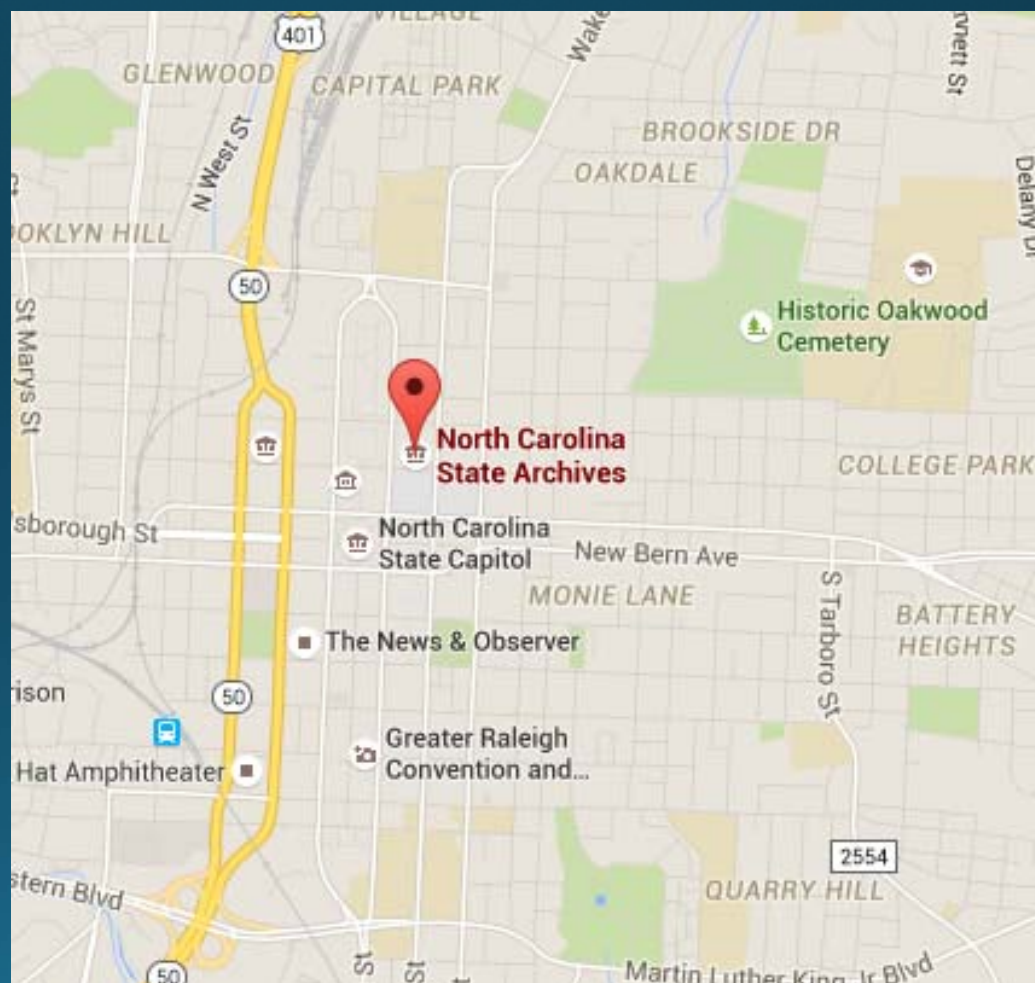
Where We've Been...

- NC processes Gov. Jim Hunt emails, 2001-2002
- EMCAP, 2007-2009
- Executive Orders 150 and 18, 2009
- Mimosa Nearpoint, 2009-2014



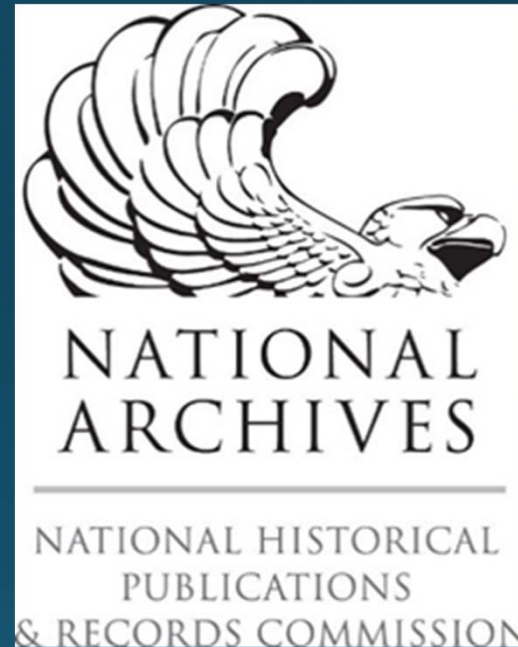
Where We Are...

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

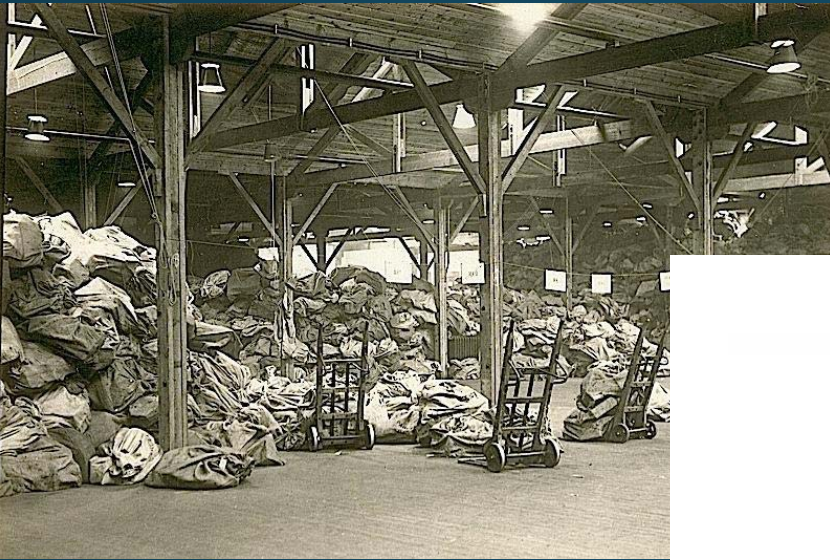


Transforming Online Mail with Embedded Semantics (TOMES)

- 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

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



Collecting Data



archives.ncdcr.gov

DIVISION OF ARCHIVES AND RECORDS
GOVERNMENT RECORDS SECTION



archives.ncdcr.gov

4615 Mail Service Center

DIVISION OF ARCHIVES AND RECORDS



archives.ncdcr.gov

4615 Mail Service Center, Raleigh NC 27699-4165

919-807-7350

DIVISION OF ARCHIVES AND RECORDS
GOVERNMENT RECORDS SECTION

Agency: _____

Emails from the individuals in this category are high-level, senior-level, or otherwise significant.

For each individual identified below, please list the email accounts used by that individual.

- The head of the agency, the president, or the equivalent.** Although the one position may be held by more than one individual, list only one individual.

NAME

Predecessors (from the list above)

NAME

ASSESSMENT OF EMAIL ACCOUNTS

Agency: _____

Please list below the individuals in the suggested roles and positions of your agency. Core functions or programs that are essential to the mission of your agency, or that serve significant populations and/or groups, or that conduct research supporting core programs, oversee outreach programs, capture the history of core programs, or are responsible for these email accounts are not already listed under categories 1 through 4.

If the emails generated by individuals listed below are already being captured by the email account of an agency position listed on the Part 1 form you already completed, including instances where necessary to list those individuals here. Your agency may not list individuals who are already listed on the Part 1 form.

For each individual identified below, please list the email accounts used by that individual.

- Staff assistants to heads of agencies and their deputies.** Important work is often carried out by special assistants to senior officials and/or their email account contacts. List only those staff assistants who support senior officials.

NAME	POSITION TITLE/ROLE	BEACON POSITION NUMBER

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:

- Major agency policies
- Formulation of rules and monitoring standards (e.g., Administrative Code)
- 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
- Direction and planning of the core program(s) of your agency
- Cooperation with external state and/or federal agencies
- Construction and real property transactions
- Major public events, such as the State Fair, First Flight Centennial, inaugurations, etc.
- History of the state of North Carolina
- Advocacy for minorities, such as Indian tribes
- Certification, commissioning, etc.
- Management of assets held in public trust for the people of North Carolina – e.g. state parks, historic sites, artifacts, archival materials, etc.
- Evaluation of rules created by other agencies, where the agency is an established part of the rule review process

NAME	POSITION TITLE/ROLE	BEACON POSITION NUMBER	ARCHIVAL SUBJECT MATTER (insert number from list above)	EMAIL ADDRESS	BEGINNING DATE FOR EMAIL COLLECTION

Plugging In

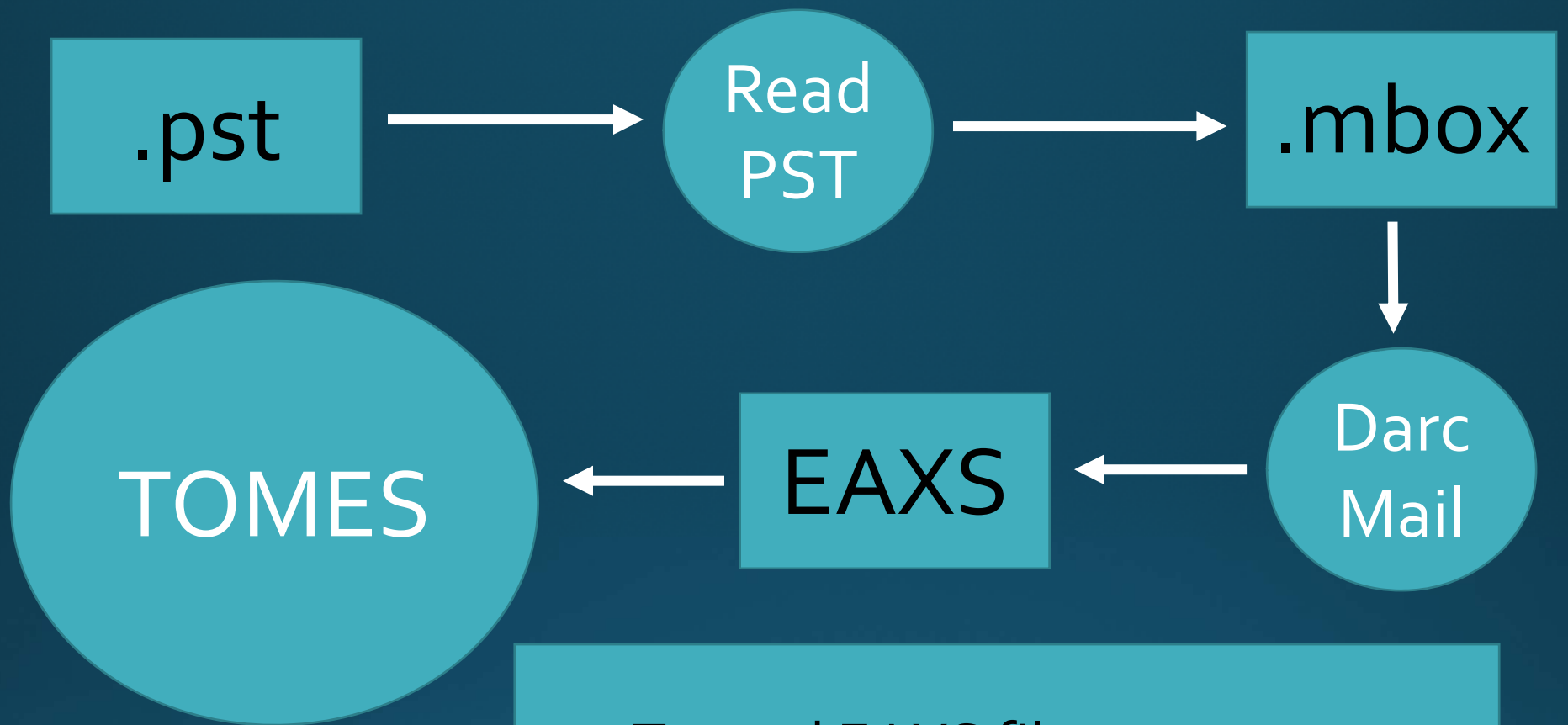
- Department of Information Technology
 - “Tagging” accounts by function
 - Facilitating the transfer of email accounts from cloud storage
- Office of State Human Resources
 - Identifying positions by position number
 - Working with DIT to “tag” accounts



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





- 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

```
<SingleBody>
  <Charset>us-ascii</Charset>
  <BodyContent>
    <Content>
      <![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.
      ]]>
    </Content>
  </BodyContent>
</SingleBody>
```

```
<SingleBody>
  <Charset>us-ascii</Charset>
  <BodyContent>
    <Content>
      <![CDATA[
        <tagged_content xmlns="http://archives.ncdcr.gov/mail-account
        Hi,

        Can the TOMES software recognize SSNs?

        For example, would it tag <tagged entity="PII.social_security
        Don't worry, I got the number from <tagged entity="ORGANIZATI

        Thanks,
        <tagged entity="PERSON" authority="stanford.edu" id="3">Fred<

        ps: I'm enjoying my vacation from <tagged entity="LOCATION" a

        </tagged_content>
      ]]>
    </Content>
  </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.
```

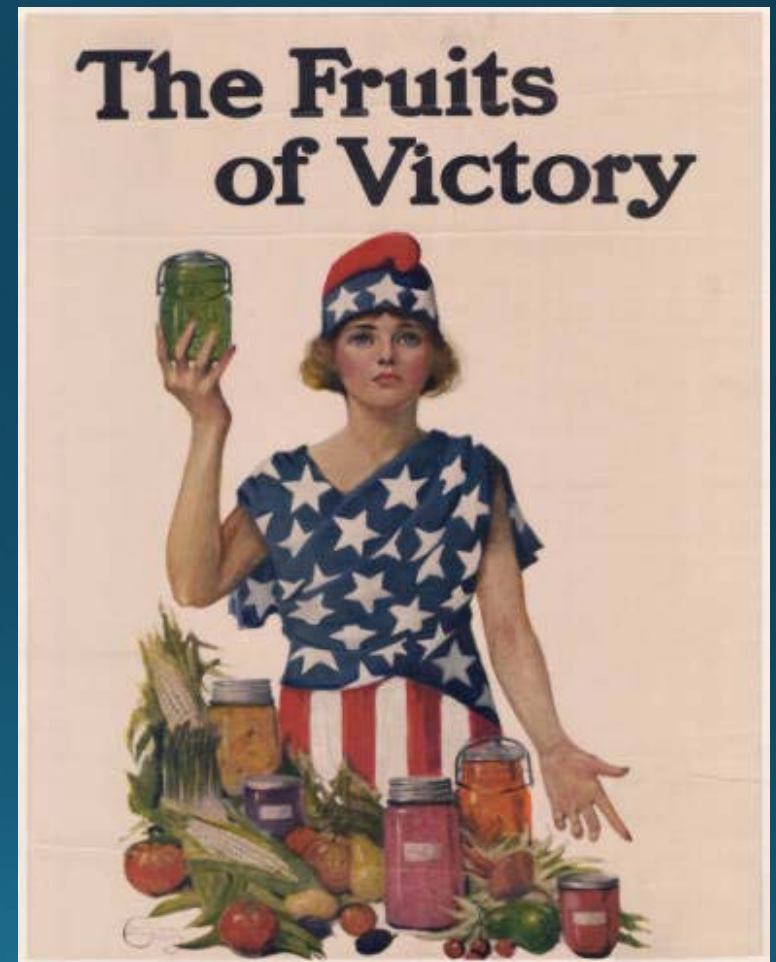
Challenges

- 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



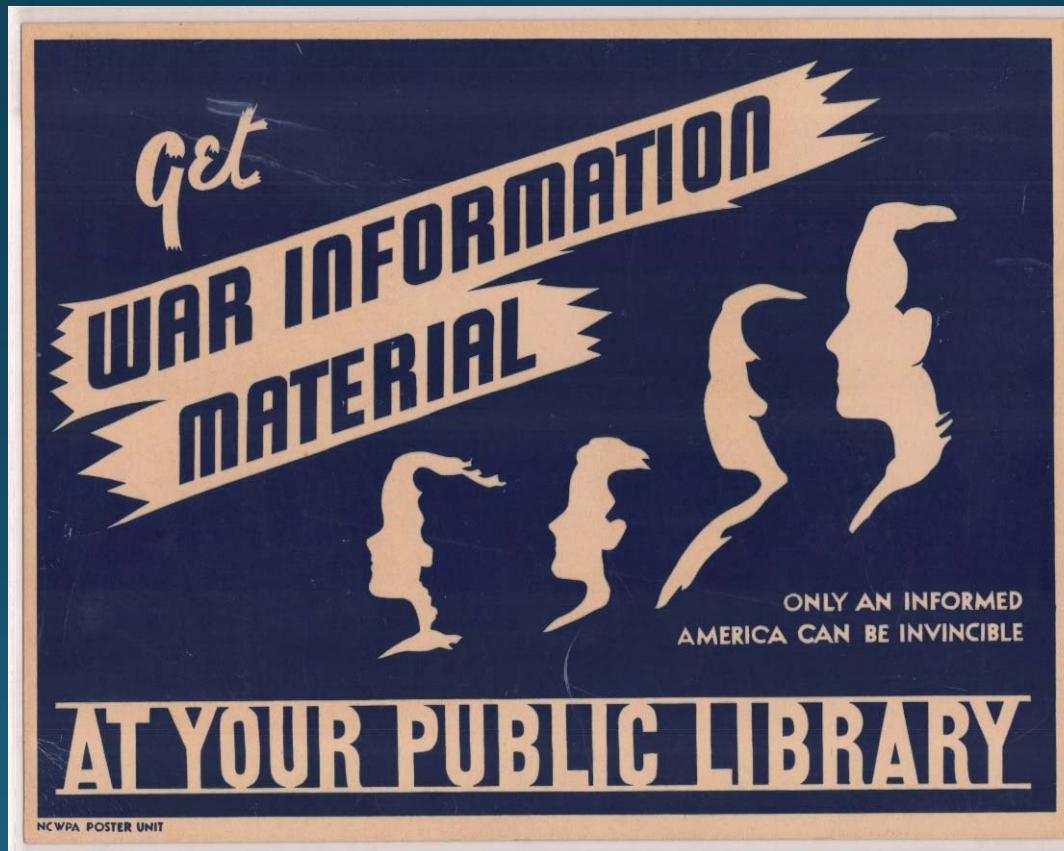
The Final Product

- 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



State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES