# Managing and Accessing Archival Email
# The TOMES Project

Camille Tyndall Watson
Digital Services Section Head
State Archives of NC

Jeremy Gibson
Systems Integration Librarian
State Archives of NC

NATIONAL ARCHIVES
NATIONAL HISTORICAL PUBLICATIONS & RECORDS COMMISSION

State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES

# Your Program for Today

1) History of SANC's email program
2) TOMES (Capstone)
3) TOMES (Tool)

# 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

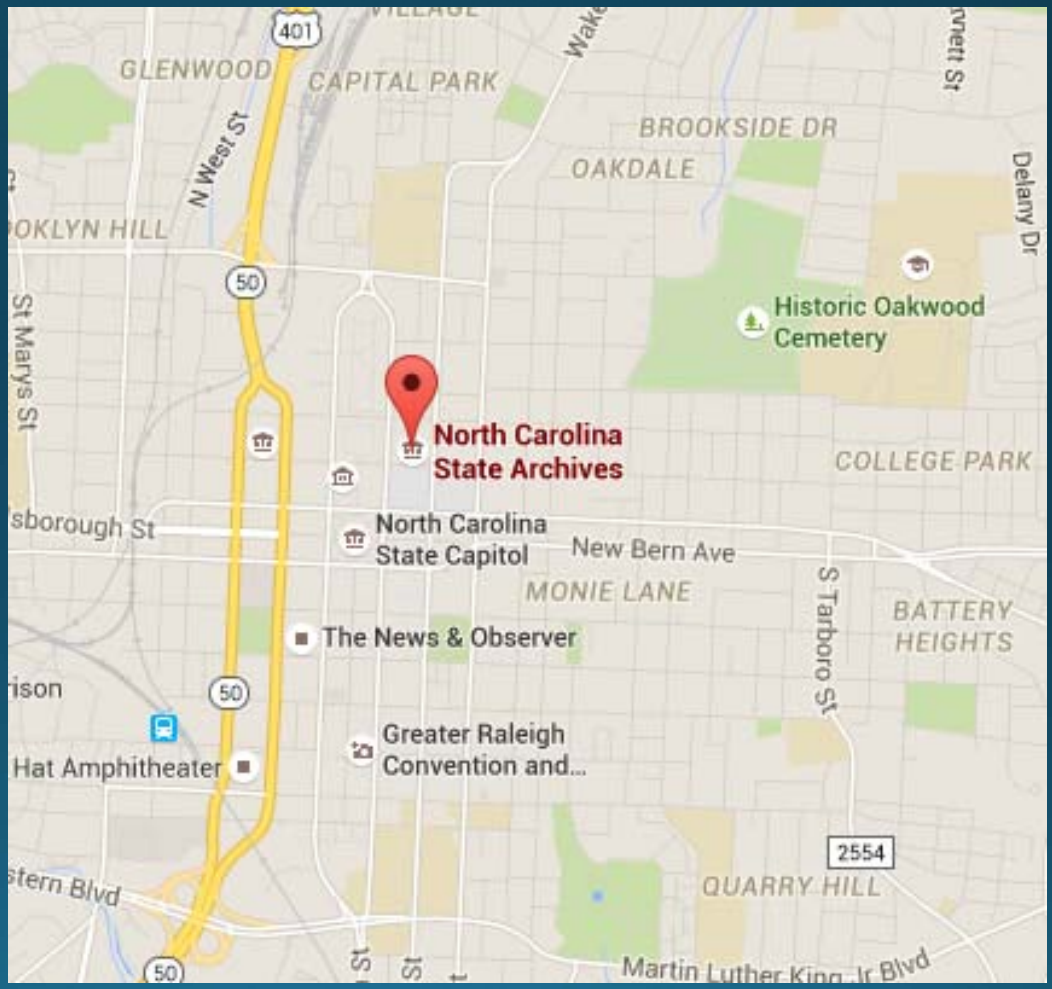



Image: Google Maps

# 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



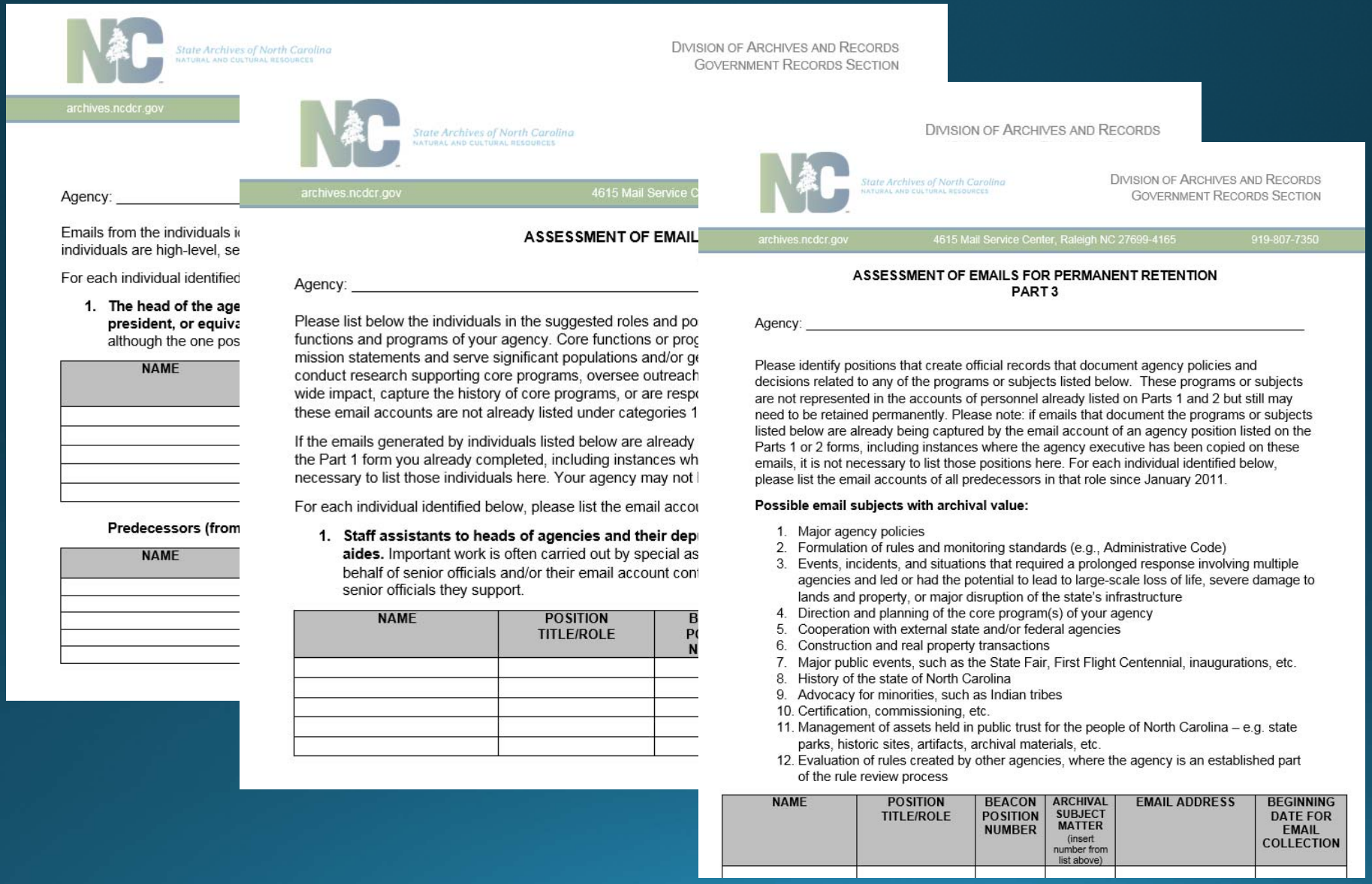

# What should we keep?


Image: 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

# Collecting Data



State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES

DIVISION OF ARCHIVES AND RECORDS
GOVERNMENT RECORDS SECTION

archives.ncdcr.gov

Agency: _____

Emails from the individuals i
individuals are high-level, se

For each individual identified

1. **The head of the age**
   **president, or equiva**
   although the one pos

| NAME |
|---|
| |
| |
| |
| |
| |

**Predecessors (from**

| NAME |
|---|
| |
| |
| |
| |
| |

State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES

archives.ncdcr.gov 4615 Mail Service C

**ASSESSMENT OF EMAIL**

Agency: _____

Please list below the individuals in the suggested roles and po
functions and programs of your agency. Core functions or prog
mission statements and serve significant populations and/or ge
conduct research supporting core programs, oversee outreach
wide impact, capture the history of core programs, or are respo
these email accounts are not already listed under categories 1

If the emails generated by individuals listed below are already
the Part 1 form you already completed, including instances wh
necessary to list those individuals here. Your agency may not

For each individual identified below, please list the email accou

1. **Staff assistants to heads of agencies and their dep**
   **aides.** Important work is often carried out by special as
   behalf of senior officials and/or their email account cont
   senior officials they support.

| NAME | POSITION TITLE/ROLE | B P N |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |

DIVISION OF ARCHIVES AND RECORDS

State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES

DIVISION OF ARCHIVES AND RECORDS
GOVERNMENT RECORDS SECTION

archives.ncdcr.gov 4615 Mail Service Center, Raleigh NC 27699-4165 919-807-7350

**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

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

# Plugging In

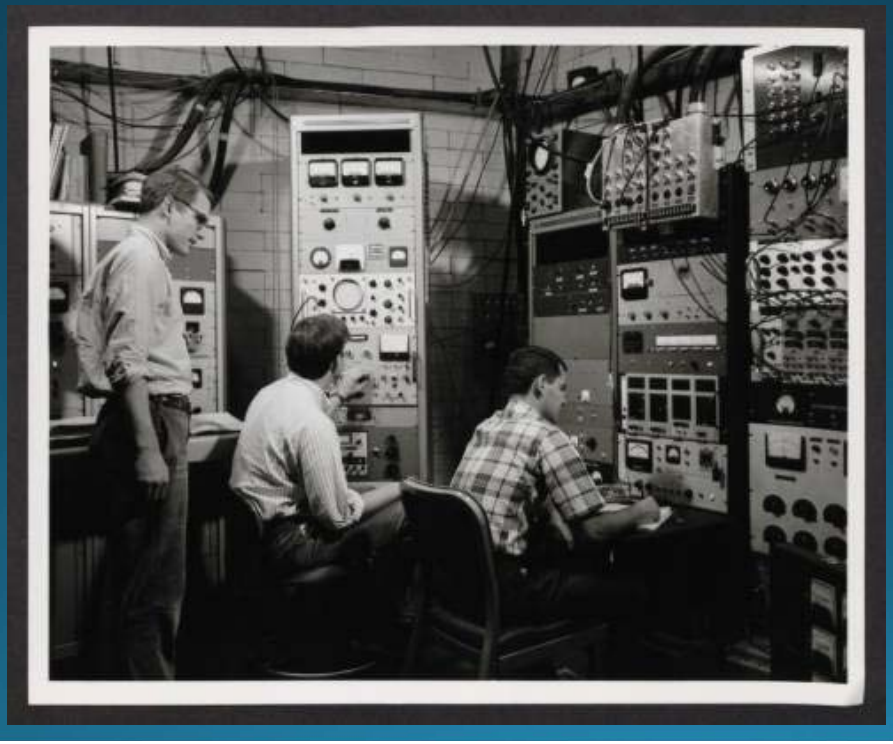
- Department of Information Technology
  - "Tagging" accounts by function
  - Facilitating the transfer of email accounts from cloud storage
- Office of State Human Resources
  - Identifying positions by position number
  - Working with DIT to "tag" accounts

# Challenges

- Capstone and Account identification
- Development of appraisal criteria
- Getting buy-in and support from agency CROs and CIOs
- Outreach and education
- Communicating needs to IT
- Figuring out who needs to be at the table

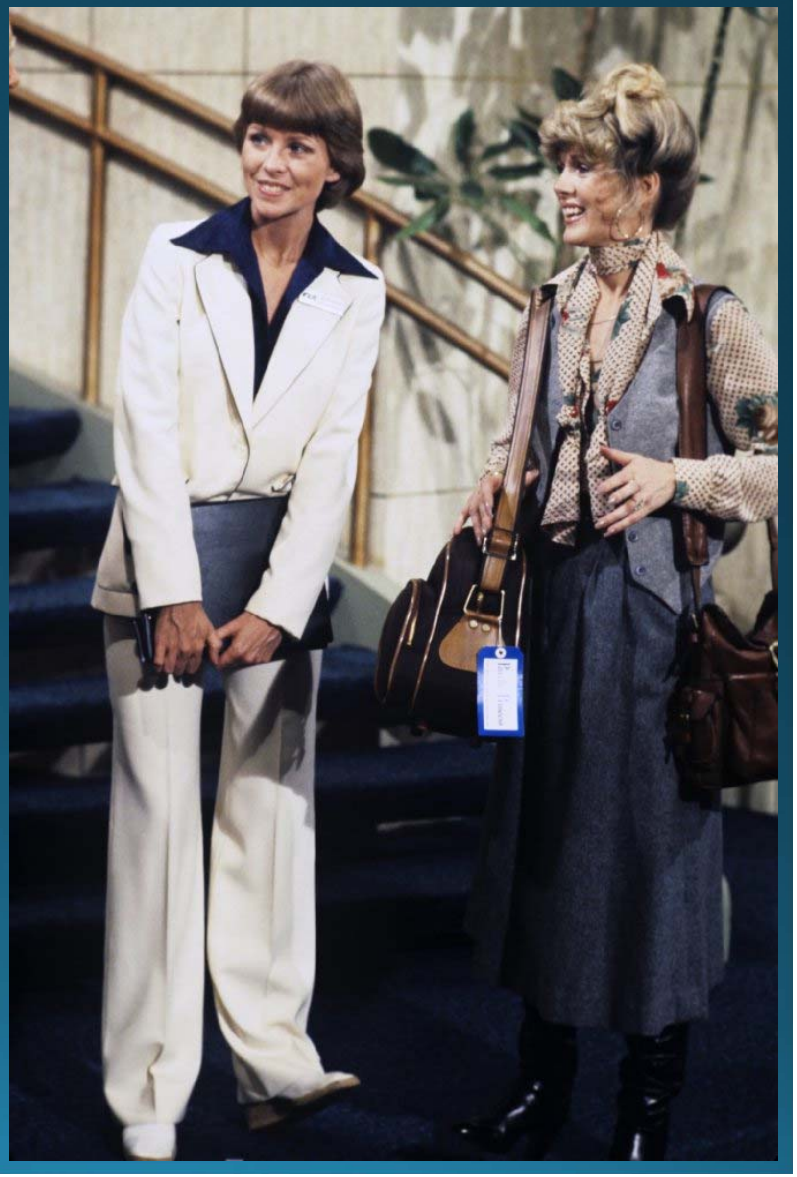
# Making It Work

- 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



# TOMES (Tool)

1) Architecture/Design
2) Original Goals
3) Modified Goals
4) V1.0



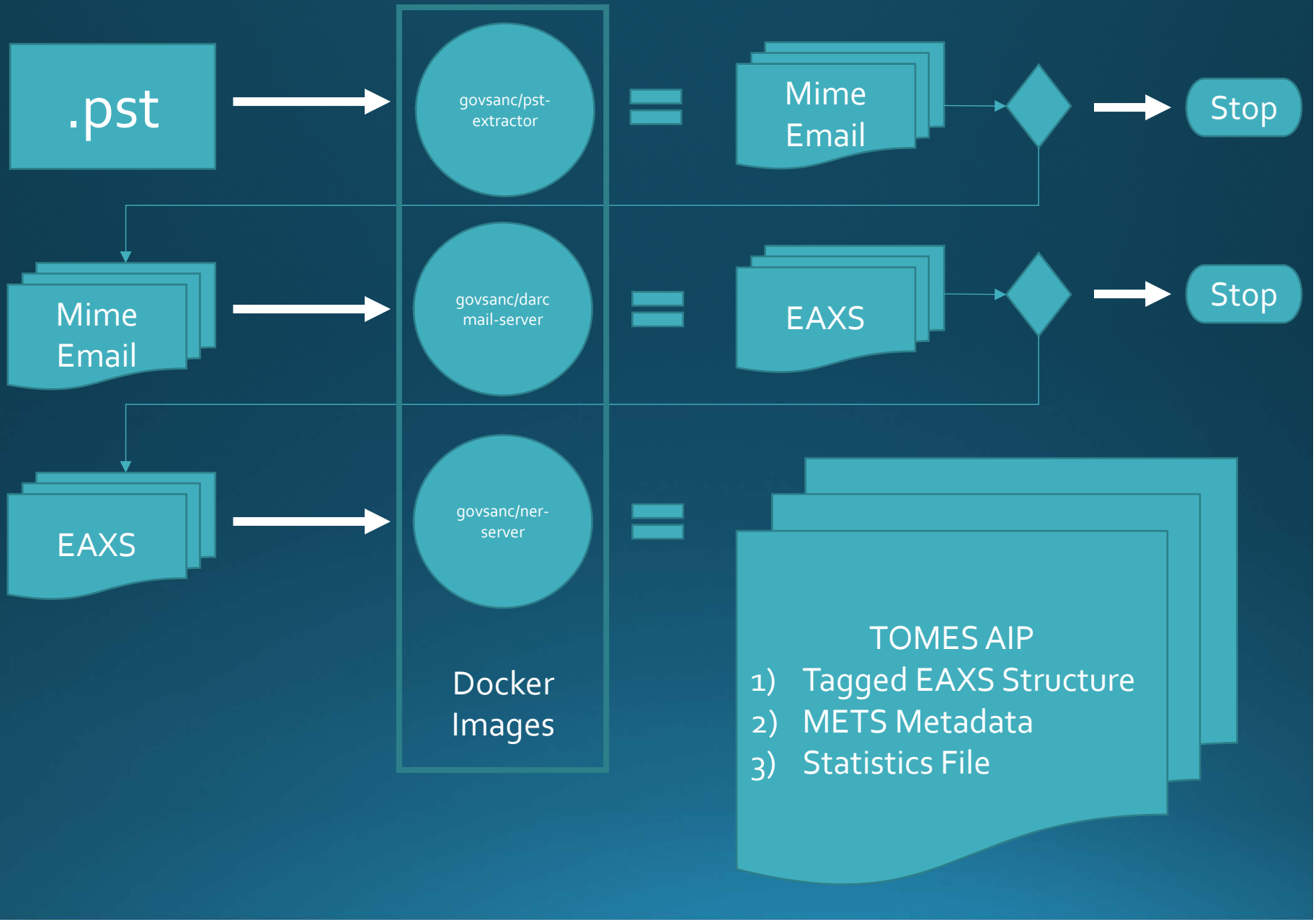
# Architecture/Design

- KISS (Keep It Simple Stupid)
- Microservices

# Architecture: KISS

- Use the best tool for the job
- Don’t reinvent the wheel
  - Readpst
  - DarcMail
  - CoreNLP

# Architecture: Microservices



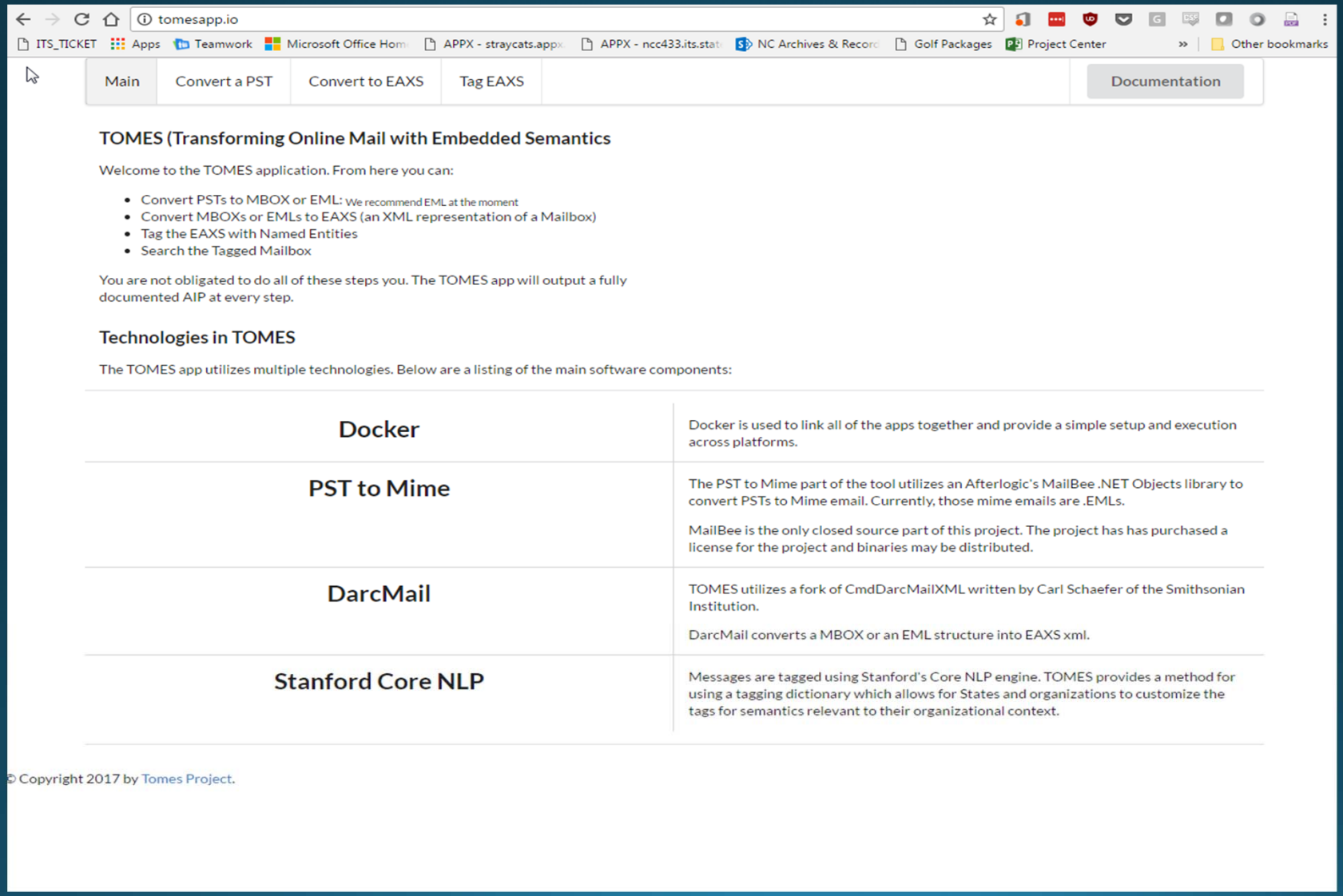
# Original Goals

- Move PSTs from black box to preservation format
- Add semantics to the text to aid in processing/access
  - Semantics customizable by institution
- Run on hardware/software available to under resourced institutions.

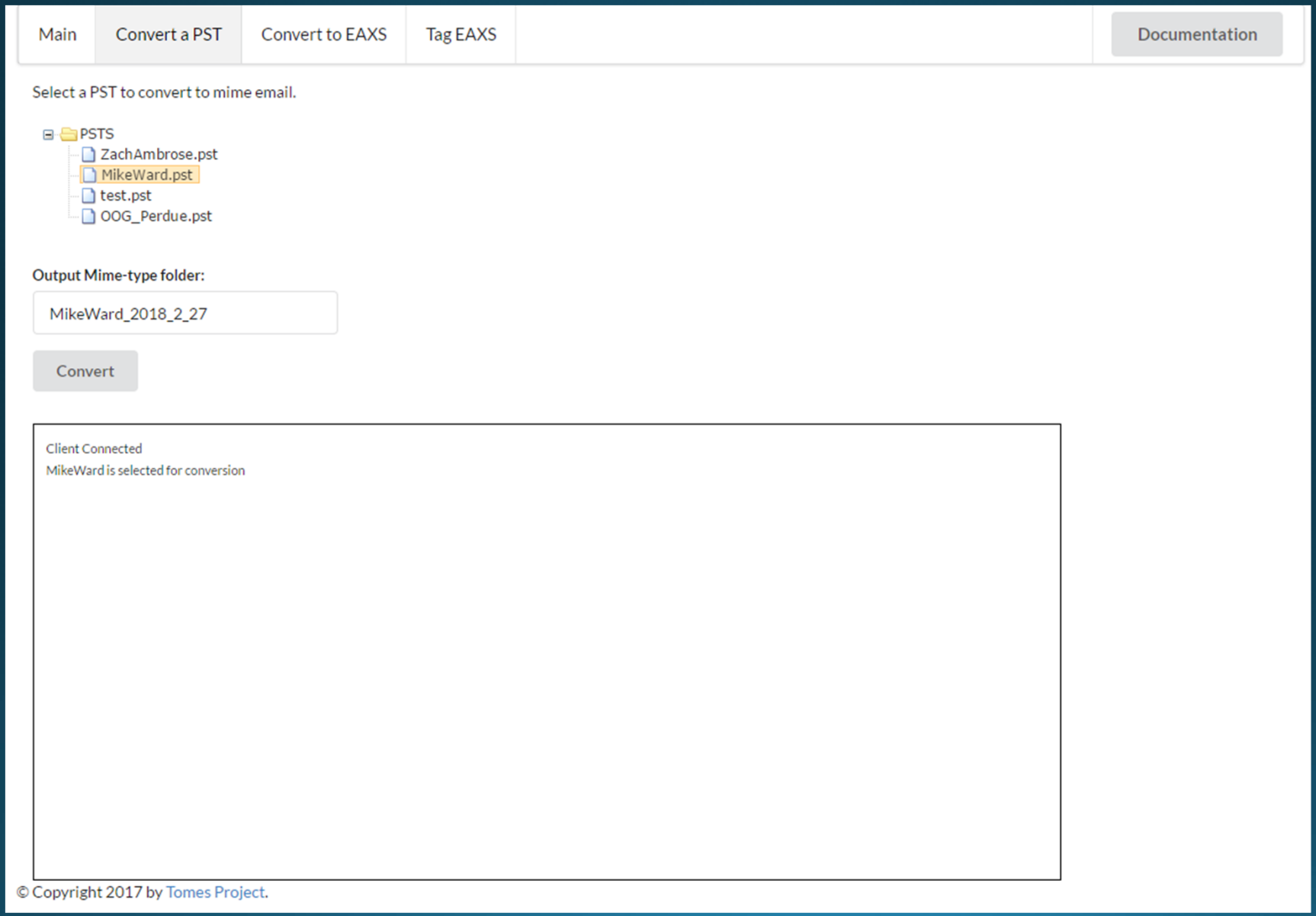
# Modified Goals

- Allow for iterative processing
- Greater focus on the PII part of the semantics
- Make the package self describing and atomic

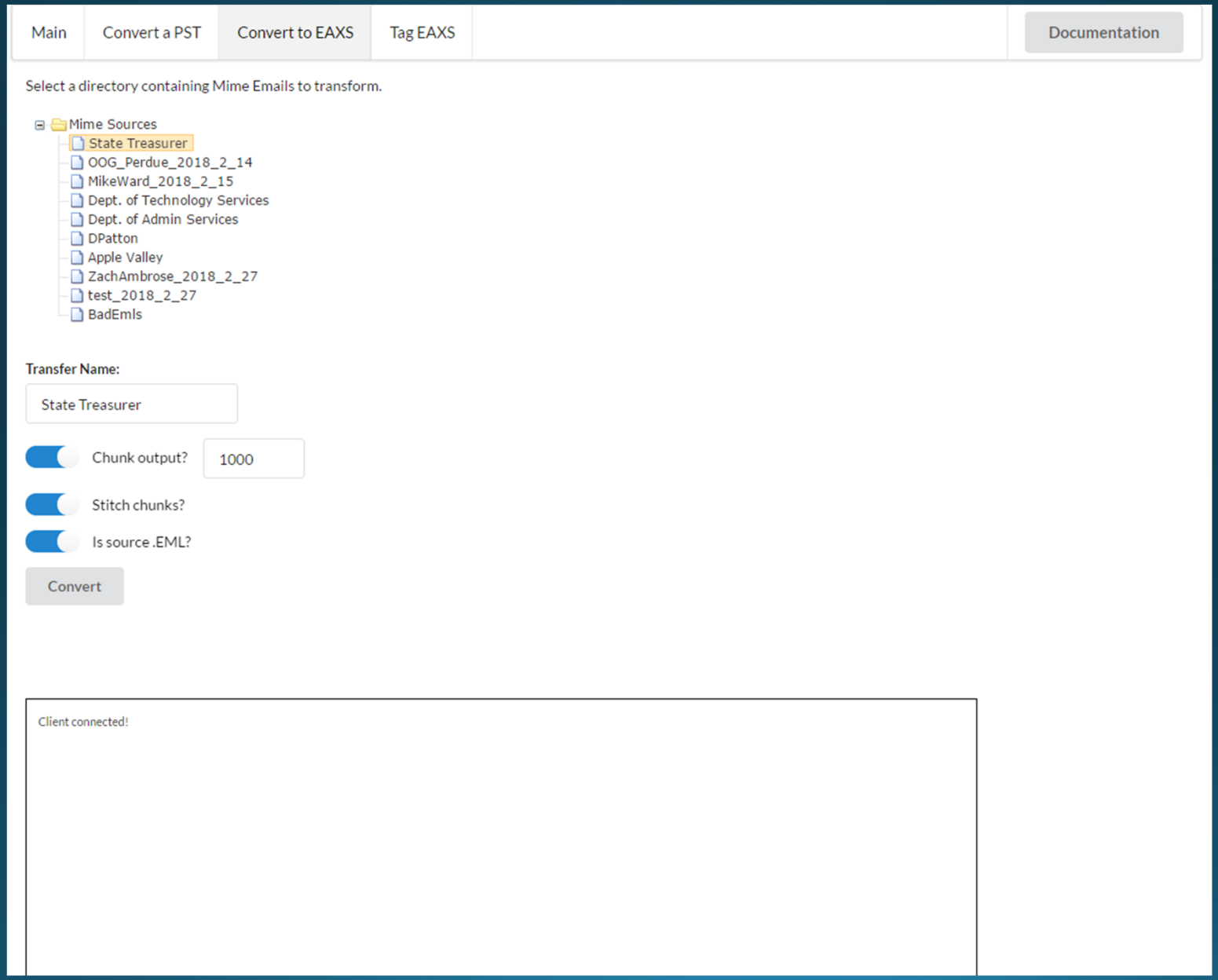
# TOMES v1.0


tomesapp.io
ITS_TICKET
Apps
Teamwork
Microsoft Office Hom
APPX - straycats.appx
APPX - ncc433.its.stat
NC Archives & Recorc
Golf Packages
Project Center
Other bookmarks
Main
Convert a PST
Convert to EAXS
Tag EAXS
Documentation
TOMES (Transforming Online Mail with Embedded Semantics
Welcome to the TOMES application. From here you can:
Convert PSTs to MBOX or EML: We recommend EML at the moment
Convert MBOXs or EMLs to EAXS (an XML representation of a Mailbox)
Tag the EAXS with Named Entities
Search the Tagged Mailbox
You are not obligated to do all of these steps you. The TOMES app will output a fully documented AIP at every step.
Technologies in TOMES
The TOMES app utilizes multiple technologies. Below are a listing of the main software components:
Docker
Docker is used to link all of the apps together and provide a simple setup and execution across platforms.
PST to Mime
The PST to Mime part of the tool utilizes an Afterlogic's MailBee .NET Objects library to convert PSTs to Mime email. Currently, those mime emails are .EMLs.
MailBee is the only closed source part of this project. The project has has purchased a license for the project and binaries may be distributed.
DarcMail
TOMES utilizes a fork of CmdDarcMailXML written by Carl Schaefer of the Smithsonian Institution.
DarcMail converts a MBOX or an EML structure into EAXS xml.
Stanford Core NLP
Messages are tagged using Stanford's Core NLP engine. TOMES provides a method for using a tagging dictionary which allows for States and organizations to customize the tags for semantics relevant to their organizational context.
© Copyright 2017 by Tomes Project.

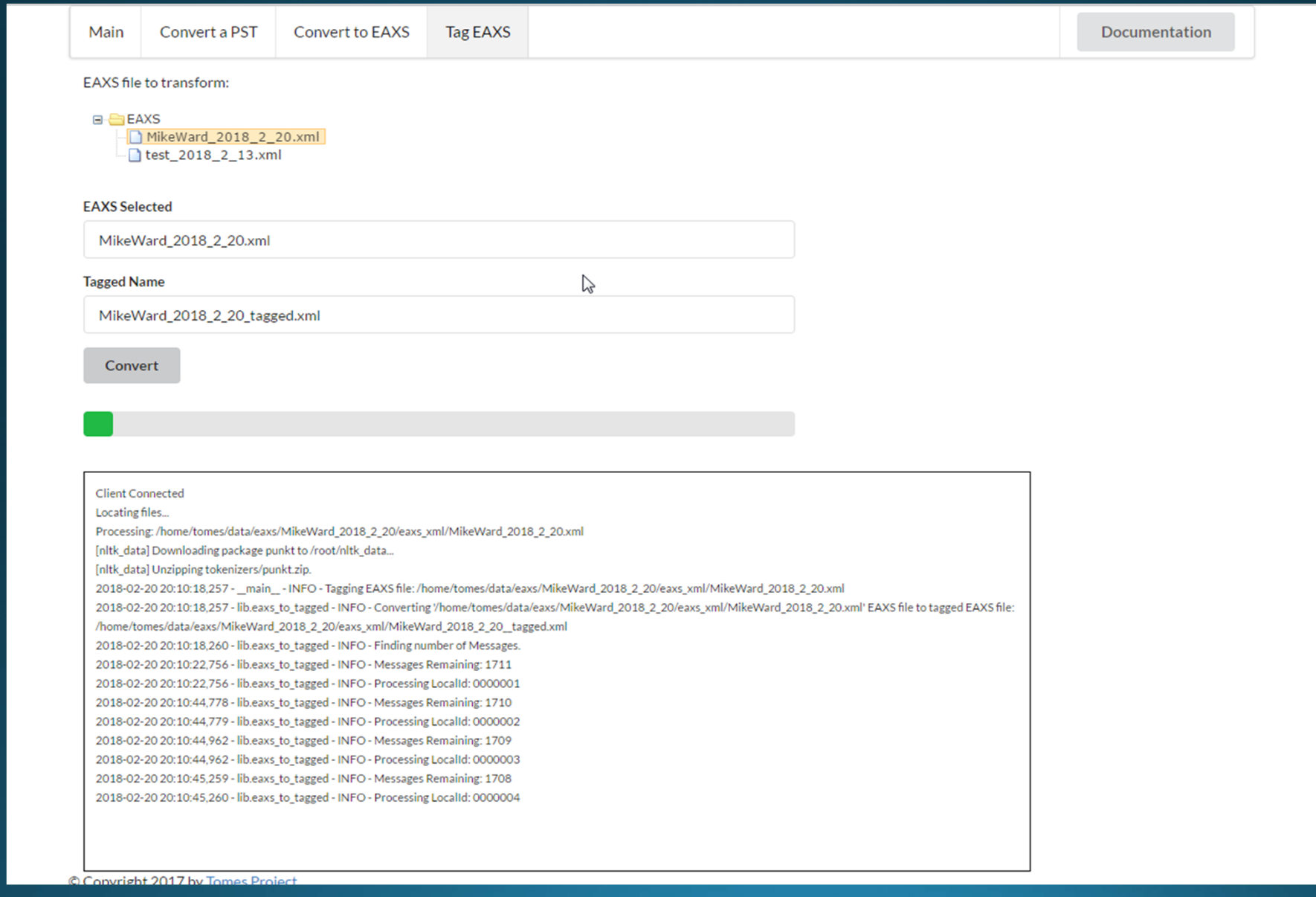
# TOMES v1.0


Main
Convert a PST
Convert to EAXS
Tag EAXS
Documentation
Select a PST to convert to mime email.
PSTS
ZachAmbrose.pst
MikeWard.pst
test.pst
OOG_Perdue.pst
Output Mime-type folder:
MikeWard_2018_2_27
Convert
Client Connected
MikeWard is selected for conversion
© Copyright 2017 by Tomes Project.

# TOMES v1.0

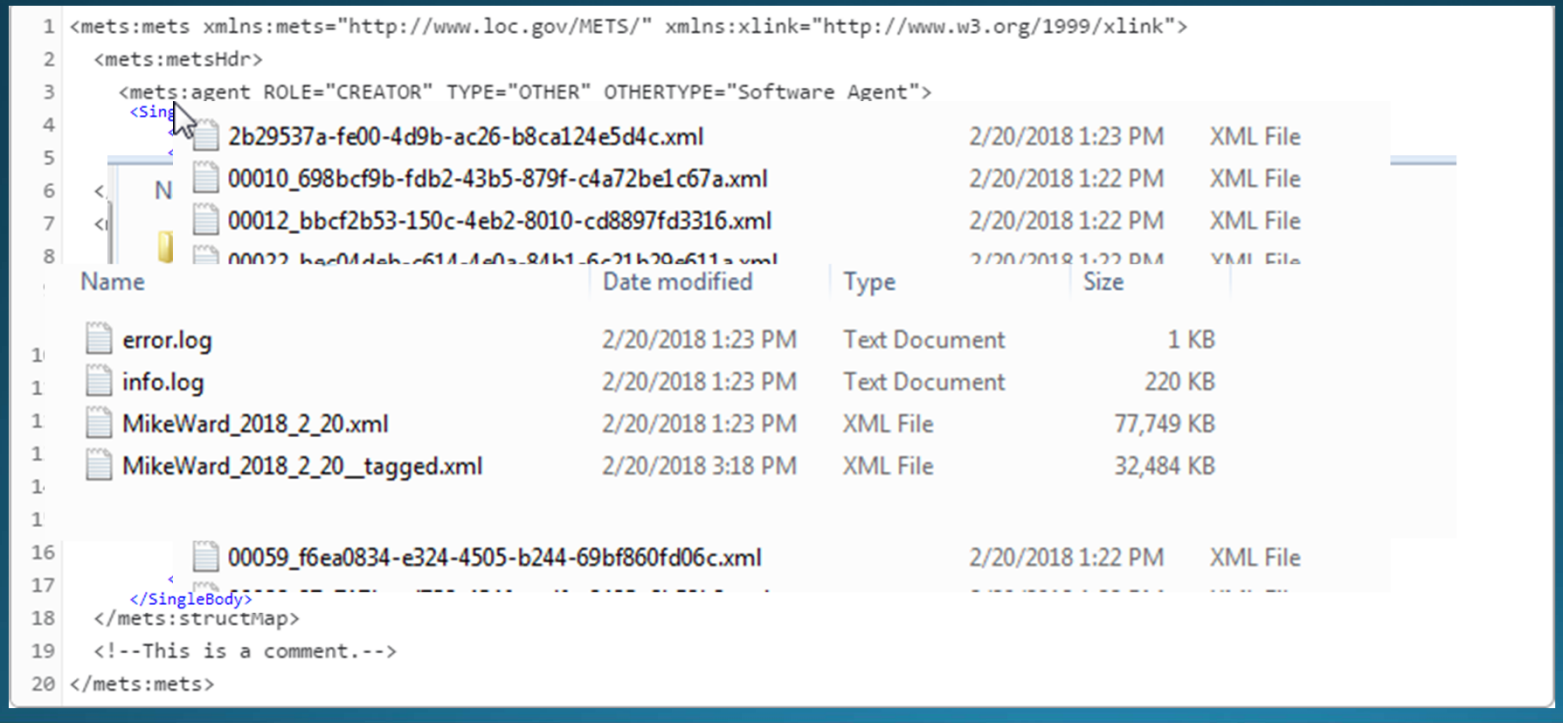

Main
Convert a PST
Convert to EAXS
Tag EAXS
Documentation
Select a directory containing Mime Emails to transform.
Mime Sources
State Treasurer
OOG_Perdue_2018_2_14
MikeWard_2018_2_15
Dept. of Technology Services
Dept. of Admin Services
DPatton
Apple Valley
ZachAmbrose_2018_2_27
test_2018_2_27
BadEmls
Transfer Name:
State Treasurer
Chunk output?
1000
Stitch chunks?
Is source .EML?
Convert
Client connected!

# TOMES v1.0


Main
Convert a PST
Convert to EAXS
Tag EAXS
Documentation
EAXS file to transform:
EAXS
MikeWard_2018_2_20.xml
test_2018_2_13.xml
EAXS Selected
MikeWard_2018_2_20.xml
Tagged Name
MikeWard_2018_2_20_tagged.xml
Convert
Client Connected
Locating files...
Processing: /home/tomes/data/eaxs/MikeWard_2018_2_20/eaxs_xml/MikeWard_2018_2_20.xml
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
2018-02-20 20:10:18,257 - __main__ - INFO - Tagging EAXS file: /home/tomes/data/eaxs/MikeWard_2018_2_20/eaxs_xml/MikeWard_2018_2_20.xml
2018-02-20 20:10:18,257 - lib.eaxs_to_tagged - INFO - Converting '/home/tomes/data/eaxs/MikeWard_2018_2_20/eaxs_xml/MikeWard_2018_2_20.xml' EAXS file to tagged EAXS file: /home/tomes/data/eaxs/MikeWard_2018_2_20/eaxs_xml/MikeWard_2018_2_20_tagged.xml
2018-02-20 20:10:18,260 - lib.eaxs_to_tagged - INFO - Finding number of Messages.
2018-02-20 20:10:22,756 - lib.eaxs_to_tagged - INFO - Messages Remaining: 1711
2018-02-20 20:10:22,756 - lib.eaxs_to_tagged - INFO - Processing LocalId: 0000001
2018-02-20 20:10:44,778 - lib.eaxs_to_tagged - INFO - Messages Remaining: 1710
2018-02-20 20:10:44,779 - lib.eaxs_to_tagged - INFO - Processing LocalId: 0000002
2018-02-20 20:10:44,962 - lib.eaxs_to_tagged - INFO - Messages Remaining: 1709
2018-02-20 20:10:44,962 - lib.eaxs_to_tagged - INFO - Processing LocalId: 0000003
2018-02-20 20:10:45,259 - lib.eaxs_to_tagged - INFO - Messages Remaining: 1708
2018-02-20 20:10:45,260 - lib.eaxs_to_tagged - INFO - Processing LocalId: 0000004
© Copyright 2017 by Tomes Project

# V1.0: The Final AIP

- Original Account file (.pst, .mbox, .eml)
- Untagged EAXS XML file
- Attachment XML file(s)
- Tagged EAXS XML file
- METs file

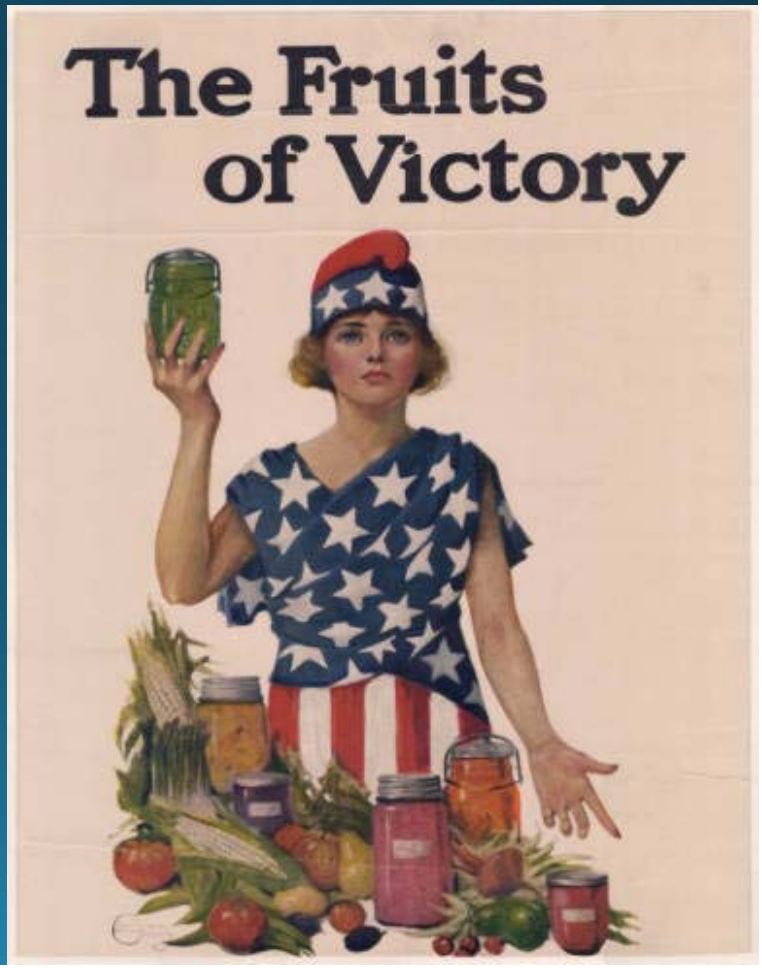


# Challenges

- Making imperfect NER useful.
  - Machine Learning
  - Active discovery and processing
- Making State specific libraries easier for non-technical users to develop and incorporate into the workflow.
- Handling of emails with bad encodings
  - Email is messy and comes from everywhere.

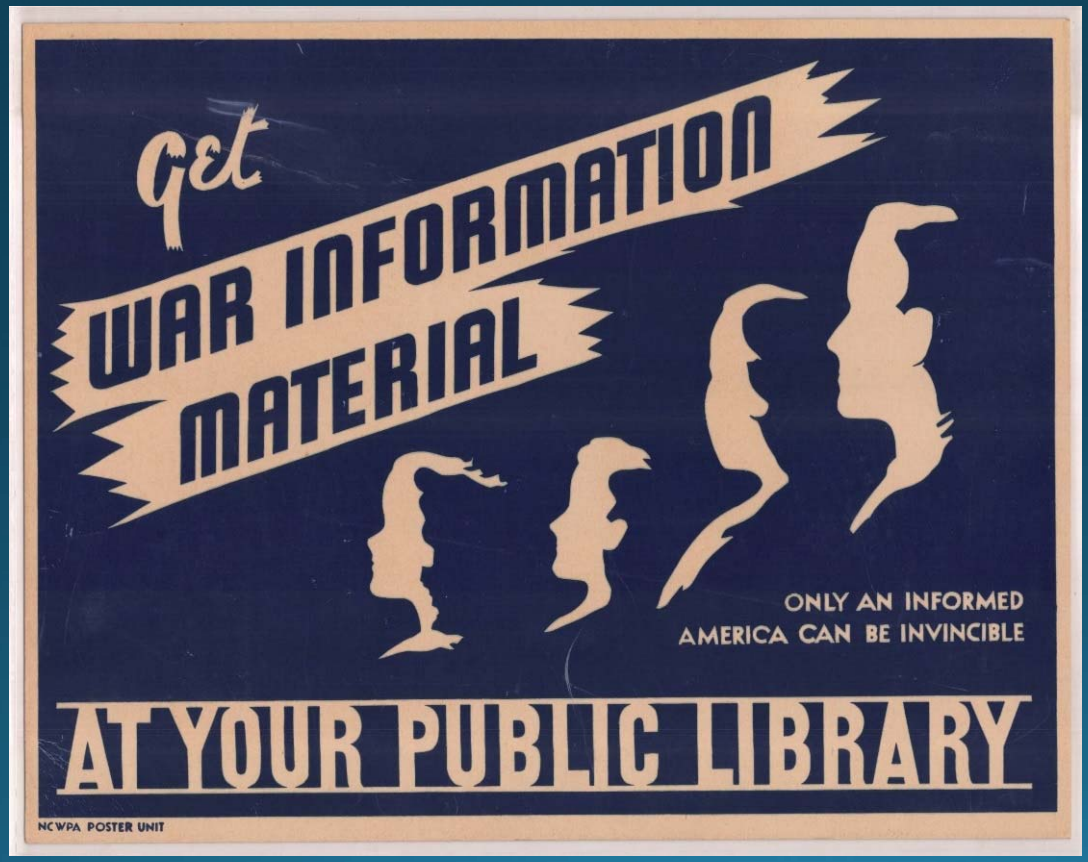
# The Final Product

- 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

Jeremy Gibson
Systems Integration Librarian
State Archives of NC
Jeremy.Gibson@ncdcr.gov
(919) 807-7356

State Archives of North Carolina
NATURAL AND CULTURAL RESOURCES