New Tools and Techniques to Analyze and Manage Email in Archives

Development Updates from the Review, Appraisal, and Triage of Mail Project

Christopher (Cal) Lee and Kam Woods
UNC Chapel Hill School of Information and Library Science
OPFCON Lightning Talk - June 9, 2020

Review, Appraisal, and Triage of Mail (RATOM)

- Funded by the Andrew W. Mellon Foundation (2019-2020)

- Developing software for selection and appraisal of email using NLP and other machine learning, including a web app to assist archivists evaluating email materials for retention and release

- Support iterative processing - information discovered during the processing workflow can support further selection, redaction or description actions

- Mapping of timestamp, entity, sensitive features and other elements across tools

- Extends workflows developed for BitCurator and TOMES

Ray Tomlinson
Implemented first email program on ARPANET. Credited with invention of first email system.
Team Members

- Cal Lee
  PI

- Antoine De Torcy
  Software Engineer

- Camille Tyndall Watson
  Co-PI

- Jamie Patrick-Burns
  Investigator

- Eliscia Kinder
  Project Manager

- Kam Woods
  Technical Lead (UNC)

- Sangeeta Desai
  Technical Lead (NC DAR)

- Caktus Group
  Software Development
RATOM tools - libratom

libratom (reusable library)

Python library to parse and analyze PST, OST, and mbox email formats

Wraps functions from libpff, Python mailbox, and spaCy (NLP)

Email message content, header, attachment extraction; entity identification and classification

Engineered to scale with core count and keep memory use flat per-core

https://www.github.com/libratom/libratom

Generating features that are:

VERIFIABLE
REPRODUCIBLE
REUSABLE

facilitates understanding changes in assessments of materials over time.
RATOM tools - Iterative Processing Interface

Assist archivists in reviewing email materials for retention and/or release.

- Import of email accounts from PSTs and entity identification via libratom
- Creation of processing accounts associated with individual email users
- Interactive review and tagging of email messages within these accounts (e.g. “record”, “non-record”, “redact”)
- Export of selected messages as EML for retention or release

https://github.com/StateArchivesOfNorthCarolina/ratom-deploy
Accounts View

Accounts associated with imports of one or more imported PST files are displayed in the main interface.

Account processing indicates Complete when all entity identification and full-text indexing has finished.
Individual Account

Selecting an account displays an infinite-scroll view of individual messages associated with that account.

Green tags indicate entity classes identified during processing.

Status dropdown allows messages to be marked for retention or redaction (also appears in individual message view).
Message View

Messages are tagged during ingest using categories associated with entities identified in the body text.

(Note: this research dataset contains prior annotations, resulting in overtagging)
Tagging and Search

Selection by classification (e.g. record vs non-record) and date range.
Audit History

Audit histories for individual messages are retained, ensuring a clear record of initial processing actions and potential changes over time.
Project info, news, and blog posts:
https://ratom.web.unc.edu/
Core library:
https://github.com/libratom/libratom
Sample Jupyter notebooks:
https://github.com/libratom/ratom-notebooks
Web app (iterative processing interface):
https://github.com/StateArchivesOfNorthCarolina/ratom-deploy

@RATOM_Project