

# New Tools and Techniques to Analyze and Manage Email in Archives

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

RATOM 

**Christopher (Cal) Lee and Kam Woods**

UNC Chapel Hill School of Information and Library Science

OPFCON Lightning Talk - June 9, 2020

VIEW THESE SLIDES AT: <https://bit.ly/ratom-opfcon-2020>



UNC  
SCHOOL OF INFORMATION  
AND LIBRARY SCIENCE



NC DEPARTMENT OF  
NATURAL AND CULTURAL RESOURCES



# 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)



**Cactus Group**  
Software Development

# RATOM tools - libratom

## libratom (reusable library)

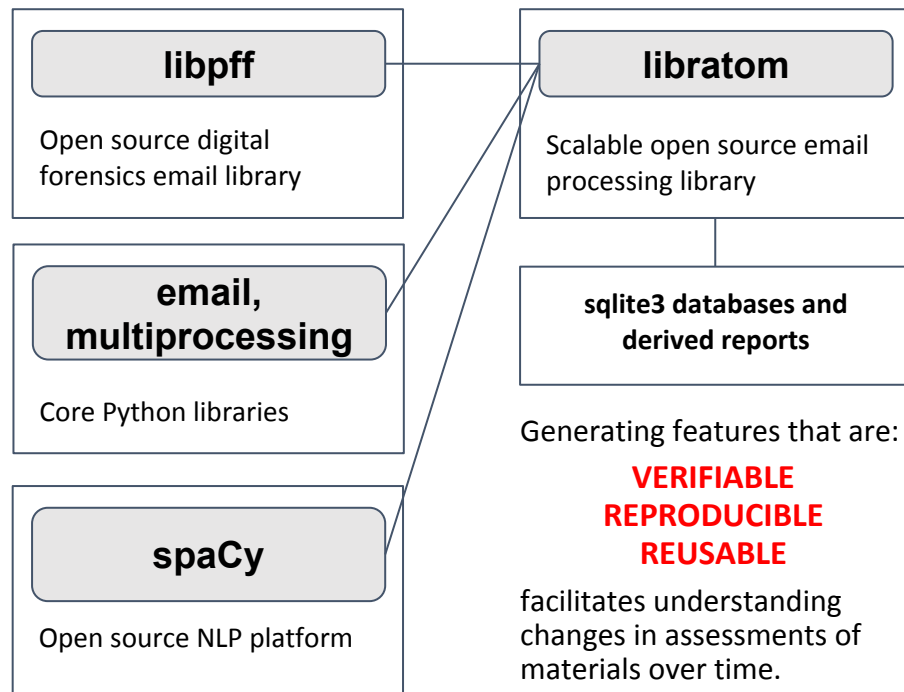
Python library to parse and analyze **PST**, **OST**, and **mbox** email formats

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

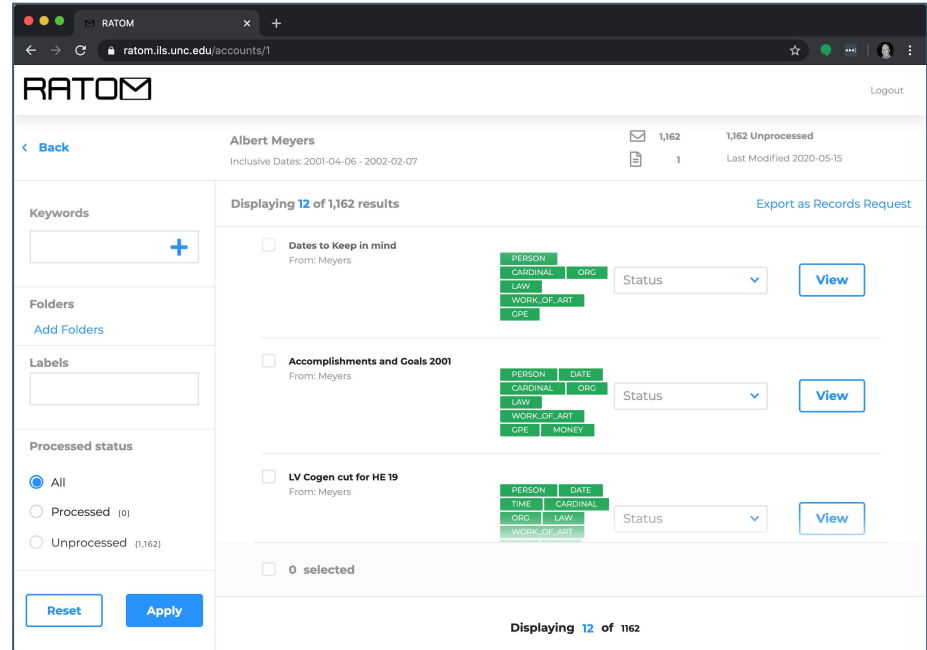
<https://www.github.com/libratom/libratom>



# 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

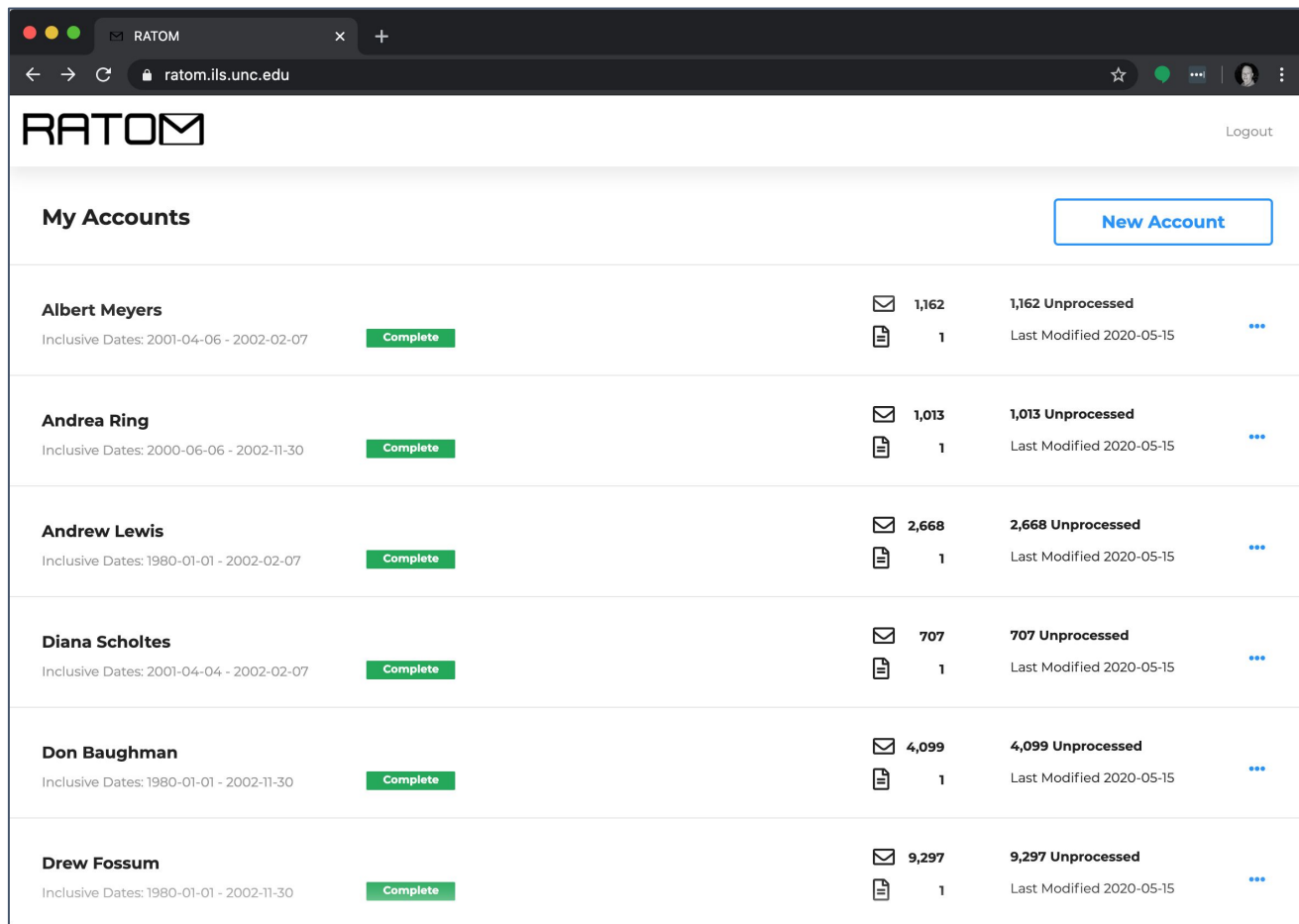


<https://github.com/StateArchivesOfNorthCarolina/ratom-deploy>

## 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.



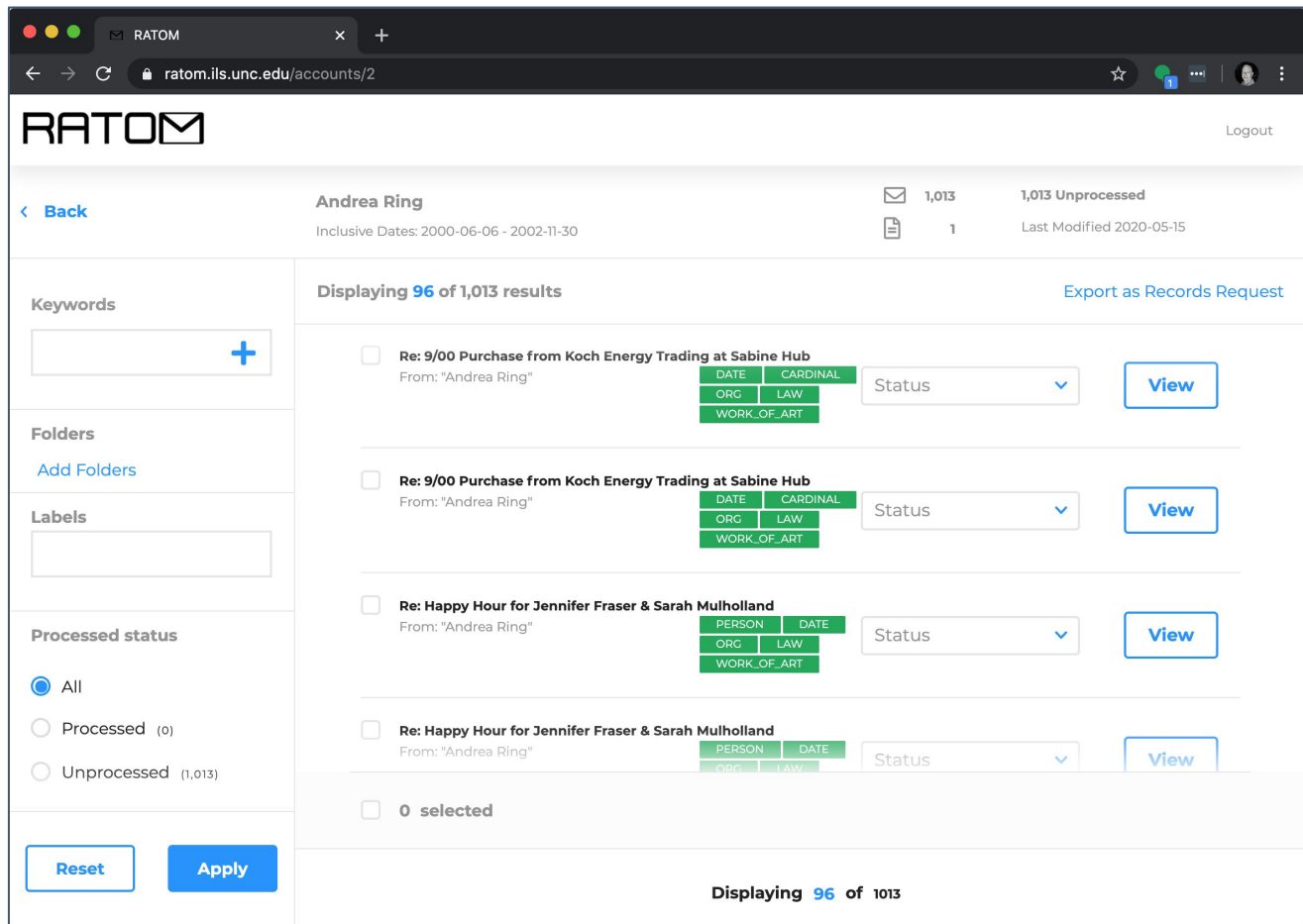
My Accounts				<a href="#">New Account</a>	
<b>Albert Meyers</b> Inclusive Dates: 2001-04-06 - 2002-02-07	Complete	1,162 1	1,162 Unprocessed Last Modified 2020-05-15	...	
<b>Andrea Ring</b> Inclusive Dates: 2000-06-06 - 2002-11-30	Complete	1,013 1	1,013 Unprocessed Last Modified 2020-05-15	...	
<b>Andrew Lewis</b> Inclusive Dates: 1980-01-01 - 2002-02-07	Complete	2,668 1	2,668 Unprocessed Last Modified 2020-05-15	...	
<b>Diana Scholtes</b> Inclusive Dates: 2001-04-04 - 2002-02-07	Complete	707 1	707 Unprocessed Last Modified 2020-05-15	...	
<b>Don Baughman</b> Inclusive Dates: 1980-01-01 - 2002-11-30	Complete	4,099 1	4,099 Unprocessed Last Modified 2020-05-15	...	
<b>Drew Fossum</b> Inclusive Dates: 1980-01-01 - 2002-11-30	Complete	9,297 1	9,297 Unprocessed Last Modified 2020-05-15	...	

# 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)

The screenshot shows a web browser window with the RATOM interface. The browser's address bar displays the URL `ratom.ils.unc.edu/accounts/2/messages/686`. The RATOM logo is in the top left, and a 'Logout' link is in the top right. Below the logo is a navigation bar with a '< Back' link, a '37 of 1013' indicator, a 'View as plain-text' checkbox (which is checked), and a 'Status' dropdown menu. The main content area displays an email with the subject 'Re: 8/00 Purchase from Koch Energy Trading at Sabine (Henry Hub) - Sitara Deal' and a timestamp of 'Sep 19, 2000 11:04 PM'. The email headers show 'To: "Michael Mousteiko"' and 'From: "Andrea Ring"'. Below the headers, a row of green tags is visible: PERSON, CARDINAL, ORG, LAW, WORK\_OF\_ART, GPE, and MONEY. A '+ Add Label' link is positioned below these tags. A breadcrumb trail reads '> Top of Personal Folders > ring-a > Andrea\_Ring\_Jun2001 > Notes Folders > Sent'. The email body is separated from the header by a dashed line labeled 'START MESSAGE BODY'. The text of the email states: 'I do not show I did any deals outside of EOL with Koch at the Sabine Hub during this time frame. Sitara deal #369260 refers to EOL deal ID 369298.' This is followed by a disclaimer block starting with '\*\*\*\*\*', mentioning 'EDRM Enron Email Data Set' and 'ZL Technologies, Inc.', and ending with another '\*\*\*\*\*'. The body is separated from the footer by a dashed line labeled 'END MESSAGE BODY'. At the bottom of the interface, a '37 of 1013' indicator is present.



# Tagging and Search

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

RATOM

Logout

Andrea Ring

Inclusive Dates: 2000-06-06 - 2002-11-30

1,013 1,013 Unprocessed

1 Last Modified 2020-05-15

Displaying 24 of 1,013 results

Export as Records Request

Record status

☒ All

☐ Open (0)

☐ Restricted (0)

☐ Needs redaction (0)

☐ Non-record (0)

Email addresses

+

From:

YYYY-MM-DD

To:

Reset Apply

☐ Re: Happy Hour During Gas Fair  
From: "Andrea Ring"

PERSON ORG  
LAW  
WORK\_OF\_ART

Status View

☐ Re: Happy Hour During Gas Fair  
From: "Andrea Ring"

PERSON ORG  
LAW  
WORK\_OF\_ART

Status View

☐ From: "Andrea Ring"

PERSON DATE  
ORG LAW  
WORK\_OF\_ART

Status View

☐ From: "Andrea Ring"

PERSON DATE  
ORG LAW

Status View

☐ 0 selected

Displaying 24 of 1013

# Audit History

Audit histories for individual messages are retained, ensuring a clear record of initial processing actions and potential changes over time.

Select message audit to change

ratom.ils.unc.edu/admin/core/messageaudit/

WELCOME, RATOM@PROTONMAIL.COM. VIEW SITE / CHANGE PASSWORD / LOG OUT

Home > Core > Message audits

Select message audit to change

ADD MESSAGE AUDIT +

Q

Search

Action:  Go 0 of 100 selected

<input type="checkbox"/>	PK	MESSAGE	PROCESSED	IS RECORD	DATE PROCESSED	UPDATED BY
<input type="checkbox"/>	22035	Re: AGC Job Posting...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22034	Revised Draft...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22033	Re: Article...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22032	Re: New Revised Draft Answer - Ignore Pr...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22031	...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22030	Lodi Storage...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22029	FW: CONFIRMATION: April 20, 2001 Executi...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22028	Houston...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22027	Re: Your Law Conference RSVP Form...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22026	Submit Your Law Conference RSVP Form...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22025	...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22024	Re: FW: Draft Transwestern Response to F...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22023	Re: Additional Needles capacity...	<div></div>	<div></div>	-	-
<input type="checkbox"/>	22022	Re: Chicago update	<div></div>	<div></div>	-	-

FILTER

By is record

All

Yes

No

By processed

All

Yes

No

By account

All

Albert Meyers

Andrea Ring

Andrew Lewis

Diana Scholtes

Don Baughman

Drew Fossum

Dutch Quigley



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

