IJDC | *Conference Paper*

Open-Source Email Curation Software Designed for Reusability

Christopher A. Lee
University of North Carolina
at Chapel Hill - SILS

Kam Woods
University of North Carolina
at Chapel Hill - SILS

Abstract

Email is more than half a century old and fills a vital role in activities across all sectors of society. However, the professional curation of email is still relatively immature. Many proprietary tools that extract and query email content operate as black boxes and cannot be easily evaluated or integrated with other digital curation software. Recently, there has been progress in the development of open-source software for email curation, but many institutions struggle to integrate these tools into digital curation workflows that usually involve other applications and systems.

We report on a project called Review, Appraisal and Triage of Mail (RATOM) that developed software for interactive review, selection and appraisal of email collections held in PST, OST, and mbox formats. These tools allow users to create, validate, and query reports and metadata generated from email collections. We have designed the software for reusability from the ground up. The software is open source and distributed as several independent modules that can be incorporated into existing and emerging workflows. Output is structured to be easily queried in support of new and emerging tasks and access scenarios. We also describe a reusable tool to simplify management of machine learning models for identifying named entities in email.

Submitted March 13, 2022 ~ *Accepted* April 22, 2022

Correspondence should be addressed to Christopher A. Lee, 216 Lenoir Drive #3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: callee@ils.unc.edu

This paper was presented at International Digital Curation Conference IDCC22, Edinburgh, 13-16th June 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v16i1.0
DOI: 10.2218/ijdc.v16i1.0

Introduction

Email is more than half a century old and has played a vital role in activities across all sectors of society. A small but essential percentage of email has continuing value, warranting care and attention of custodians, including (but not limited to) creators, records managers, archivists, as well as other professionals in galleries, libraries, archives and museums (GLAMs). However, curation of email is still relatively immature.

While various proprietary tools can extract and query email content, they are black boxes and do not afford easy evaluation or integration and data exchange with other software. Recent years have seen progress in the development of open-source software for the curation of email, but many institutions struggle with integration of tools into their overall digital curation workflows that usually involve several other applications and systems. Investigation of workflows demonstrates the vital role of hands-offs between systems and tools (Chassanoff et al, 2020).

Developing Reusable Open-Source Email Curation Software

The Review, Appraisal, and Triage of Mail (RATOM) project was a two-year partnership between the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill and the State Archives of North Carolina (SANC). Funded by the Andrew W. Mellon Foundation, RATOM developed and tested technologies to support the review and processing of email in collecting institutions. Development focused on locating messages with retention value, identifying instances of potentially sensitive information, and accurately tagging features corresponding to real-world entities such as persons, places, organizations, and events. We implemented software to scan email package files (including PST, OST, and mbox) and record and export content, metadata, and derived features such as entities identified using natural language processing (NLP) in a simple SQLite database that can be queried as part of various workflows.

In addition to the core processing tool for email package formats, the project developed a web-based tool to review email collections for retention and release. This interface allows users to import email from PST, OST, and MBOX formats; analyze message content to identify entities and other features of interest; and create a set of processing accounts associated with each ingest. Individual messages can be reviewed and tagged (for example, as “record”, “non-record”, or “requires redaction”), and message subsets can be selected for export in EML format for public release.

The RATOM tools are designed to minimize the effort required to run computationally complex email analysis tasks. For example, RATOM provides a standalone tool to replace the default named-entity recognition (NER) model with a model for a different language, a custom model, or a multi-language model. Some existing GLAM access systems incorporate NLP to describe the contents of collections, but this technology is often tightly coupled to the platform being used or is applied strictly to file types that tend to share common structures and metadata.

Email contains encoded text, markup, and attachments, but importantly also structured metadata in the header that can be used to cue identification of persons and organizations and describe their relationships. Identifying named entities, relationships, and other features of interest by processing open text from heterogeneous collections of files (such as those extracted from a disk image) is inherently "noisier," as the extracted text will often contain patterns of features (such as persons, places, and organizations) common to a wide range of devices and production environments (e.g. documentation of system files). RATOM tools provide a mechanism by which cross-format search procedures can be easily implemented – exposing header metadata in database entries where they are explicitly linked to entities identified in open text.

Related Work

Curation Guidance and Case Studies

In 2012, archivists at the University of Manchester reported on processing the email correspondence of Carcanet Publishing. iv They found that none of the tested tools met their needs, and the Carcanet correspondence remains embargoed on their servers. They also found that they needed significant amounts of storage space for working with email, along with technical knowledge that their staff did not have. Their recommendations included funding for further practical projects to address email curation.

Staff at the Library of Virginia (LVA) describe processing the email of Governor Tim Kaine (Bromley et al, 2015). LVA completed their project with an open, searchable email collection available through Virginia Memory.[1] The authors note the project revealed an "archival catch-22." LVA is mandated to provide access to government records, and the Kaine emails were in high demand during the 2016 election cycle. Online access is desirable, but government emails often contain confidential and restricted materials. This necessitated item-level processing of the emails, which they judged as being unsustainable due to funding, staffing, and forecast growth of electronic records. In their conclusions, they call for sustainable solutions for archival electronic materials.

Many other informal case studies are available online in the proceedings of archives symposia and blog posts. The blog posts are usually high level, such as the Society of American Archivist's Electronic Records Section blog post on the experiences of an archivist at Texas A&M University.[2] Symposia proceedings, such as those published by the U.S. National Archives and

[1] http://www.virginiamemory.com/collections/kaine/

[2] https://saaers.wordpress.com/2017/09/05/adventures-in-email-wrangling-tamu-ccs-epadd-story/

Records Administration (NARA) from their Government Email Symposium[3] provide significant detail regarding the issues and potential solutions for email curation, but also reveal that many LAMs are creating ad hoc workarounds to manage email.

The literature on email curation is frequently composed of high-level guidance and technical overviews, including those by the Digital Curation Centre (DCC), the Digital Preservation Coalition (DPC), NARA, and the Library of Congress (LoC). The majority of reports in this body of work call for continued community development, advocacy, and tool development to address the growing issue of archiving email.

An important contribution to this literature is the 2018 final report of the Task Force on Technical Approaches for Email Archives, sponsored by the Andrew W. Mellon Foundation and the Digital Preservation Coalition (CLIR, 2018). The report concludes with set categories of recommendations: (1) community development and advocacy, and (2) tool support, testing, and development. The latter includes several specific recommendations; developing better tools for sensitivity review - for example, to identify materials containing private or individually identifying information; and filling "the gaps in current workflows and ensure better interoperability between tools." The report of the Task Force also includes appendices that summarize tools and projects that have targeted email curation.

Tools For Curation of Email

Practical developments in email curation have grown substantially over the past decade, as have calls to accelerate related research (Prom, 2011). The Email Collection and Preservation (EMCAP) project[4], funded by the National Historical Publications and Records Commission (NHPRC), was a partnership between the State Archives of North Carolina (SANC), Pennsylvania State Archives and Kentucky Department for Libraries and Archives. The project developed the EAXS XML schema and methods for collecting email from an archive folder within a user's inbox.

ePADD (email Processing, Appraisal, Discovery, and Delivery) has been developed by Stanford University's Special Collections and University Archives since 2013. The software supports a set of appraisal, ingest, processing, and access functions for email collections. The ePADD software uses a customized Named Entity Recognition engine to identify correspondents within email. A web publication from the ePADD development group notes: "Not satisfied with other open source NER engines, including the Stanford NER and Apache OpenNLP, the ePADD development team created their own engine … to help identify and disambiguate correspondents within the corpus … [and] ensures persons that occur within the email archive who are also correspondents are weighted more heavily in this ranking." ePADD remains in active development at the time of writing, and a project to integrate a new set of preservation functionality into the software was recently funded.[5]

[3] https://www.archives.gov/nhprc/projects/electronic-records/email

[4] https://siarchives.si.edu/what-we-do/digital-curation/email-preservation-cerp

[5] https://library.stanford.edu/projects/epadd/about/eabcc-phase-4

The Transforming Online Mail with Embedded Semantics (TOMES) project[6] was a project at SANC that began in 2015. The project includes documentation on email appraisal and records management, using NARA's Capstone approach, as well as open source code and documentation for the TOMES tool. The tool uses microservices to convert PST, MBOX, and EML email account files into the EAXS XML schema[7] for secure preservation, and then use natural language processing to identify named entities, PII, and confidential material.

Email Datasets and Machine Learning

LAMs often have large and diverse collections and limited human resources, and careful applications of ML techniques can benefit processing workflows by reducing the time required to triage materials and automating bulk classification jobs that would otherwise be intractable. One of the challenges is that generating training data can be labor-intensive. In order for an ML model to classify an email message as a record (for example, an official correspondence within the preservation scope) or a non-record (anything outside the scope of preservation), tens of thousands of messages correctly annotated by a human archivist might be required for training.

Active learning (AL) is a process in which the software tries to prioritize instances for human review that are most likely to inform the underlying model. While this can improve performance, it still requires either a large amount of training data or a significant number of human expert judgements. Yue Wang and his colleagues have demonstrated a "novel interactive learning algorithm that is capable of directly acquiring domain knowledge from human experts by allowing them to articulate the evidence that leads to their sense tagging decisions (e.g., the presence of indicative words in the context that suggest the sense of the word). This knowledge is then applied in subsequent learning processes to help the algorithm achieve desirable performance with fewer iterations" (Wang et al, 2018). While they applied this approach specifically to word sense disambiguation in medical records, such interactive machine learning based on multiple forms of human input holds great promise for LAM email applications.

Another promising domain for machine learning is review for sensitivity. Email collections often contain personal identifiers, discussions of sensitive subjects, or other information that may be subject to restriction or redaction. Beginning in October 2018, the National Library of Scotland, with support from Arts and Humanities Research Council (AHRC) though the Scottish Graduate School for Arts and Humanities, Information Studies at the University of Glasgow, began a study to "use innovative methods for handling sensitive information, focusing on compliance with legal obligations (e.g. data protection)" and "investigate broader concerns, such as cost and data ethics of incorporating AI in data handling."[8]

[6] https://www.ncdcr.gov/resources/records-management/tomes
[7] https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs
[8] https://www.gla.ac.uk/colleges/arts/graduateschool/fundingopportunities/aistudentship

Appraisal and Selection of Digital Materials

Workflows and software developed for the curation of email must be attentive to LAM's policies and criteria for selection and appraisal. As the scale and format diversity of digital documents has grown, so has the importance of selection and appraisal of materials in archives.

One of the first efforts to apply software to archival appraisal was Ann Gilliland's 1995 dissertation, in which she elicited knowledge from domain experts and then attempted to develop an expert system (Gilliland-Swetland, 2016). Her findings suggest that software to support appraisal should allow LAM professionals to make individual decisions based on iterative feedback, rather than attempting to replace the human decision-maker with software.

In 2010, Harvey and Thompson discussed the prospects of using software for appraisal and re-appraisal, but they did not offer any follow-up implementation of these ideas (Harvey and Thompson, 2010). Recent case law has validated the use of "predictive coding," or text classification based on NLP rather than simple string matches, to identify subsets of records that warrant attention (Doug Smith, 2013). The National Archives of the UK has carried out an exploratory investigation of such approaches (National Archives of the UK, 2016). Several state governments in Australia have carried out similar investigations. xxviii

A small set of projects have specifically investigated the use of software to select email. Vinh-Doyle reports on exploratory efforts to use EnCase, a commercial digital forensics software suite, to identify email of continuing value (William P. Vinh-Doyle, 2018). This work identified some interesting factors for consideration but did not establish a set of methods or tools for use by other institutions. The Illinois State Archives, in partnership with the University of Illinois and with funding from the NHPRC, has attempted to use predictive coding to identify and provide appropriate access to the email messages of state agencies, based on NARA's Capstone Email approach.[9] Other authors have reported on efforts to carry out appraisal on email collections using a combination of software-supported and manual approaches. This includes the Library of Virginia's processing of email from Governor Tim Kaine's administration discussed earlier (Christman and Page, 2010). Cocciolo also conducted a case study that involved manual application of an email selection rubric (Anthony Cocciolo, 2016).

The RATOM Toolset

For the RATOM project we developed a selection of tools using open source software libraries to identify and extract the contents of email backup formats, perform NLP analysis on message content, and record content, metadata, and derived features in a self-contained database.

[9] https://www.uillinois.edu/cio/services/rims/about_rims/projects/processing_capstone_email_using_predictive_coding/

One group in our team focused on core email format analysis, data extraction, and feature identification tasks. While some institutions receive email as discrete document files, or directly from a remote server using a network application programming interface (API), many collections have or continue to receive complete email accounts retained in backup formats such as PST, OST, mbox, and EML. Such files may also be discovered on devices such as laptops, desktop computers, and decommissioned servers. Open source tools to accurately parse each of these formats are an important starting point for institutions acquiring and facilitating access to email over time. One goal for this team was to ensure that the tool could reliably extract relevant content and metadata from these formats. Another was to create machine-readable and human-readable reports describing potentially private and sensitive information within those collections. The team prioritized output formats and data structures designed to support data sharing, and simplify the creation of mechanisms to normalize generation and ingest of email derived metadata and identified features among systems and tools.

A second group in our team focused on development tasks targeting improved processing outcomes and workflow feasibility for institutions working with large email collections. First, the team translated and updated the iterative processing approaches used in the TOMES project to a web application. In this new application, information discovered at various points in the processing workflow is surfaced to support further selection, redaction or description actions. The web interface was also designed to tag, describe and export email items approved for release. Finally, the team investigated machine learning applications to support automated identification of characteristics associated with materials deemed to be public records, and materials that require redaction or closure in various datasets.

The complexity of various email curation tasks (preservation activities, information organization procedures, and access methods) is directly affected by how features are identified and classified. We have combined automated NER with an interface that allows human investigators to annotate individual email messages with tags to surface records of interest, highlight items that might require further investigation, and warn of potentially sensitive information. This can allow digital curation professionals to combine the advantages of bulk processing with the expertise of a human reviewer.

RATOM's development focused on needs conveyed to the project team by collecting institutions preparing email collections for preservation and public access. The tools were tested iteratively during the development period by both teams on both internal and publicly available email sets, and early builds of the tool were provided to our advisory panel for additional feedback.

Core Library

Our team developed a new Python library and associated command-line tools to support processing, analyzing, and producing reports from large collections of email containing files in PST, OST, and mbox email formats. The primary command line interface to this library (invoked simply by typing `ratom`) was designed with a number of performance-based and common sense requirements in mind:

1. The tool is multithreaded, and will use all available processing cores on the machine for text analysis by default, or a number specified by the user on demand.
2. The tool will accept as input a single file, multiple files, a directory, multiple directories, or any combination thereof, and will recurse down through any provided directory structure, processing only files identified as PST, OST, or mbox.
3. The tool outputs to a well defined SQLite database, which can be queried and transformed using a wide range of desktop and web-based tools.
4. The tool uses a default language model provided by spaCy (en_core_web_sm) for named entity recognition when no model is specified by the user; other models may easily be selected.
5. The user may specify other pre-trained models for any of the (21 at the time of writing) languages currently supported by spaCy, or provide their own. Any model specified by the user will automatically be downloaded and installed when running the tool.
6. The tool records which versions of supporting libraries and languages models are used for a particular task, so that the task may be replicated in the future even when a library or model has been superseded by a newer version.
7. The tool provides real-time feedback on completion and estimated time remaining as it is processing a collection of materials.

The core library (libratom)[10] is engineered for extensibility and ease of maintenance. It depends on libpff, an open source forensic email discovery library for processing PST and OST files, along with the Python mail library to process mbox. Libratom uses SQLAlchemy for SQL / ORM handling, and the underlying data model and schema are expressed in SQLite output generated by the associated CLI tool. The current schema focuses on initial triage and assessment; in addition to extracted email content and metadata, it provides a map of 18 unique entity types to their locations (message and file) within a corpus.

The database schema (see Figure 1) includes the elements required to link each email message to a particular source file, individual header elements to a particular message, every identified instance of an entity to a particular location within a given message body, and each attachment to a given message. This relatively compact schema captures enough context from the originating file to support a wide range of queries. For example, with the following query:

```
select count(*), label_ from entity group by label_ order by count(*) DESC;
```

we can generate a list showing how many times a particular entity type

[10] https://github.com/libratom/libratom

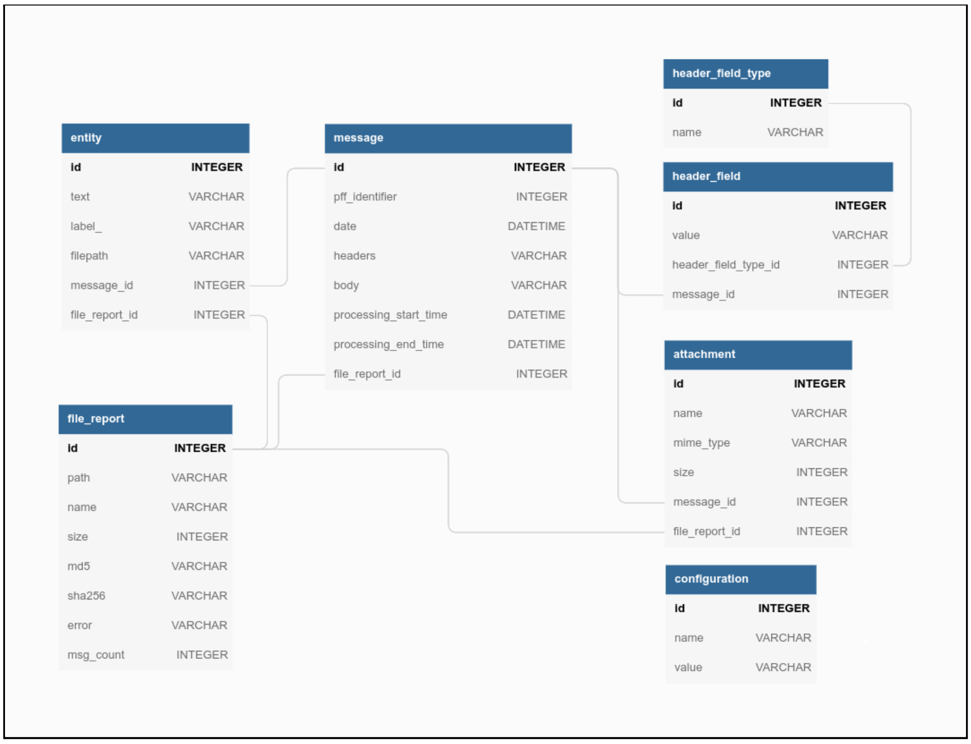
shows up in a collection, in descending order of count. Even simple queries like this can support basic insight into the common modes of communication within an email set, and may provide utility to researchers. This schema supports queries that may also give additional leverage to professionals processing collections. For example, the following query:

```
select count(*), name from attachment group by name having count(*) > 50
order by count(*) DESC;
```

returns all instances of attachments with identical names appearing within the source email file(s) more than 50 times, and a refinement of that query:

```
select count(*), name, size from attachment group by name, size having
count(*) > 50 order by count(*) DESC;
```

returns groupings of those instances where both the name and size of the attachments are identical. This is a simplified example, but demonstrates a reasonable case that queries which can be executed quickly over the database may be used to narrow the search space for certain tasks early on in a workflow (in this case, before hashes have been computed for the attachments).



header_field_type	
id	INTEGER
name	VARCHAR

entity	
id	INTEGER
text	VARCHAR
label_	VARCHAR
filepath	VARCHAR
message_id	INTEGER
file_report_id	INTEGER

message	
id	INTEGER
pff_identifier	INTEGER
date	DATETIME
headers	VARCHAR
body	VARCHAR
processing_start_time	DATETIME
processing_end_time	DATETIME
file_report_id	INTEGER

header_field	
id	INTEGER
value	VARCHAR
header_field_type_id	INTEGER
message_id	INTEGER

attachment	
id	INTEGER
name	VARCHAR
mime_type	VARCHAR
size	INTEGER
message_id	INTEGER
file_report_id	INTEGER

file_report	
id	INTEGER
path	VARCHAR
name	VARCHAR
size	INTEGER
md5	VARCHAR
sha256	VARCHAR
error	VARCHAR
msg_count	INTEGER

configuration	
id	INTEGER
name	VARCHAR
value	VARCHAR

Figure 1. Database scheme used by libratom for extracted email content, metadata and entities identified.

Tool performance is a significant concern when working with large collections. While content (individual messages and attachments) can be efficiently extracted from even very large (50GB+) individual PST files with libpff - effectively limited by disk speed - the named entity recognition task when processing individual message bodies is computationally intensive. The libratom code defaults to using all available cores when processing message bodies and headers, distributing clusters of messages to each core until the source is exhausted. Using a 32-core desktop processor (AMD Threadripper 3970X) with RAM usage capped at 64GB (2GB limit for each spaCy thread), libratom can process the entire EDRM v1.3 ENRON (54GB, 191 files, 758,341 messages) corpus in just over 1 hour. This yields an approximately 4.2GB SQLite database file when extracting both messages and entities (disregarding attachment content in this task).

Libratom is distributed with the previously described CLI utility (`ratom`) with the expectation that a significant segment of the intended audience are not software developers or dedicated IT staff, but are comfortable working with structured data output. However, it is also designed to be integrated into programmatic workflows. We have provided a set of notebooks to demonstrate how some of the tasks supported by the `ratom` tool can be replicated and customized in a modest amount of Python code.[11]

Iterative Processing Interface

The RATOM project has also developed a web application to assist archivists in reviewing email materials for retention and/or release. In its current iteration, the web application focuses on the workflow needs of institutions receiving regular email backups from external organizations and preparing them for public records requests.[12] The application is designed to support iterative processing workflows, allowing processing archivists to isolate and analyze high priority materials within a collection, tag materials according to specific organizational and legal requirements (for example, marking documents that require redaction), revisit materials that have been previously or partially processed, and maintain an audit history of actions that have been performed by specific authorized personnel to date.

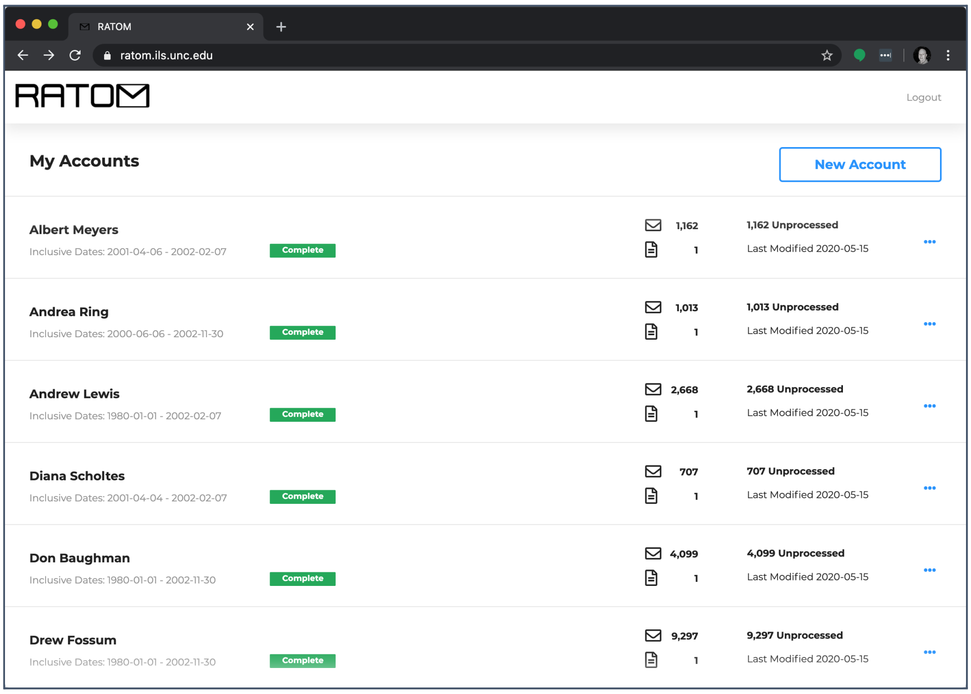
Import of email accounts from PSTs, along with NER, is handled via libratom. Accounts associated with imports of one or more imported PST files are displayed in the main interface. New accounts are displayed immediately in the interface when created, and the application monitors the server-side processing of email sources. Account processing indicates "Complete" when all NER and full-text indexing has finished.

This web application allows users to view all accounts loaded into the interface. Multiple PST files can be associated with an individual account. The interface provides a view of inclusive dates, number of messages (including number processed), number of PST files associated with the account, and last modified date on one screen. Accounts associated with imports of one or more imported PST files are displayed in the main interface. An individual account entry will not indicate "Complete" until all entity identification and full-text indexing has finished. New accounts may also be added from this view; the user may select one or more PST files

[11] https://github.com/libratom/ratom-notebooks

[12] https://github.com/StateArchivesOfNorthCarolina/ratom-web

from a shared storage location set by an administrator, and start a batch NER and indexing scan that will continue to run on the server instance even if the user leaves the page.



My Accounts

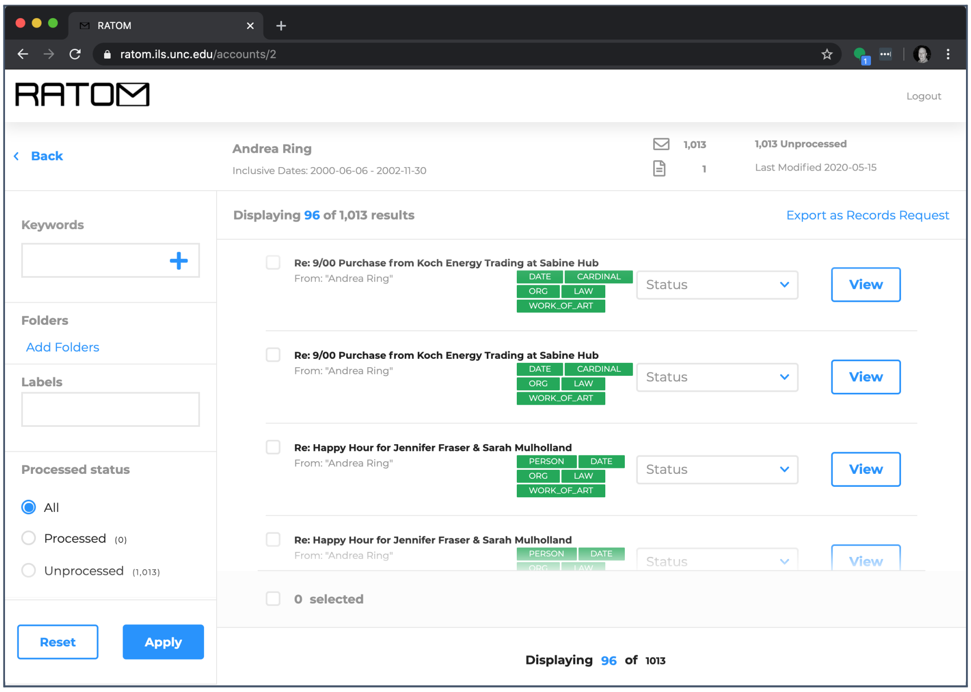
Albert Meyers Inclusive Dates: 2001-04-06 - 2002-02-07	Complete	✉ 1,162 🗎 1	1,162 Unprocessed Last Modified 2020-05-15	•••
Andrea Ring Inclusive Dates: 2000-06-06 - 2002-11-30	Complete	✉ 1,013 🗎 1	1,013 Unprocessed Last Modified 2020-05-15	•••
Andrew Lewis Inclusive Dates: 1980-01-01 - 2002-02-07	Complete	✉ 2,668 🗎 1	2,668 Unprocessed Last Modified 2020-05-15	•••
Diana Scholtes Inclusive Dates: 2001-04-04 - 2002-02-07	Complete	✉ 707 🗎 1	707 Unprocessed Last Modified 2020-05-15	•••
Don Baughman Inclusive Dates: 1980-01-01 - 2002-11-30	Complete	✉ 4,099 🗎 1	4,099 Unprocessed Last Modified 2020-05-15	•••
Drew Fossum Inclusive Dates: 1980-01-01 - 2002-11-30	Complete	✉ 9,297 🗎 1	9,297 Unprocessed Last Modified 2020-05-15	•••

Figure 2. Accounts view in the RATOM web application.

The Accounts View can be seen in Figure 2. 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. New accounts may also be added from this view; the user may select one or more PST files from a shared storage location set by an administrator, and start a batch entity identification and indexing scan that will continue to run on the server instance even if the user leaves the page.

In the Individual Account view (Figure 3), selecting an account displays a scrolling view of individual messages associated with that account. Green tags indicate entity classes identified during processing. Status drop down allows messages to be marked for retention or redaction (also appears in individual message view). The view can be filtered to show only those items that are processed or unprocessed.

The Message View (Figure 4) can be accessed from this page. Messages are tagged during ingest using categories associated with entities identified in the body text. HTML markup is filtered out of the default view to prioritize written message body content. This supports rapid review of the automatically generated label categories, determination of message body content that contains sensitive or otherwise restricted information, and tagging the message as appropriate.



Andrea Ring
Inclusive Dates: 2000-06-06 - 2002-11-30
1,013 | 1,013 Unprocessed
1 | Last Modified 2020-05-15

Displaying 96 of 1,013 results

□ Re: 9/00 Purchase from Koch Energy Trading at Sabine Hub
From: "Andrea Ring"
DATE CARDINAL ORG LAW WORK_OF_ART

□ Re: 9/00 Purchase from Koch Energy Trading at Sabine Hub
From: "Andrea Ring"
DATE CARDINAL ORG LAW WORK_OF_ART

□ Re: Happy Hour for Jennifer Fraser & Sarah Mulholland
From: "Andrea Ring"
PERSON DATE ORG LAW WORK_OF_ART

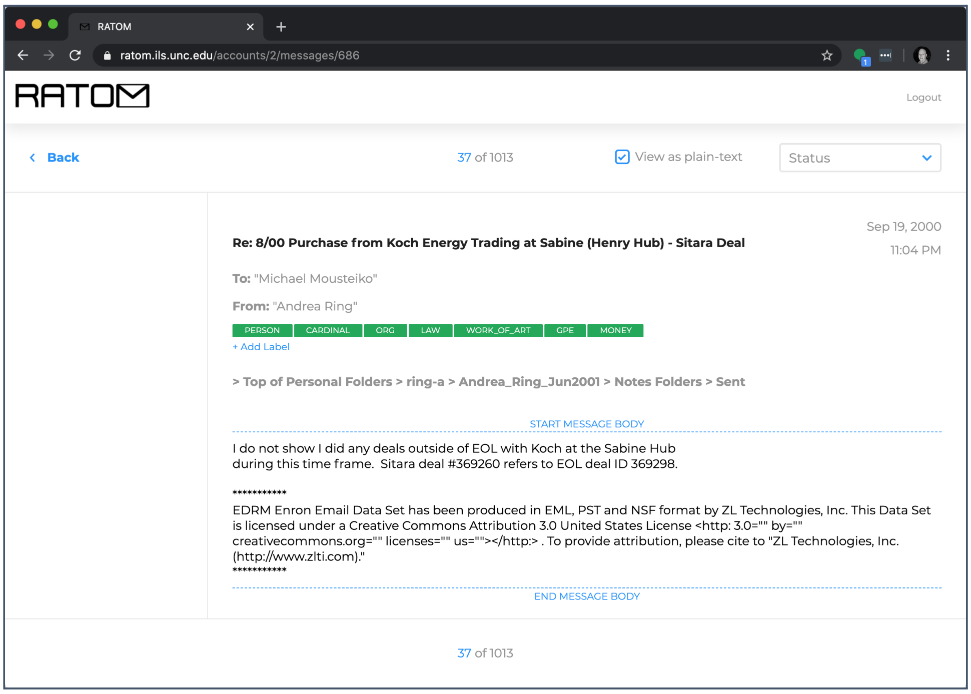
□ Re: Happy Hour for Jennifer Fraser & Sarah Mulholland
From: "Andrea Ring"
PERSON DATE ORG LAW

Processed status
● All
○ Processed (0)
○ Unprocessed (1,013)

□ 0 selected

Displaying 96 of 1013

Figure 3. Individual account view in the RATOM web application.



37 of 1013 ☑ View as plain-text

Re: 8/00 Purchase from Koch Energy Trading at Sabine (Henry Hub) - Sitara Deal
Sep 19, 2000
11:04 PM

To: "Michael Mousteiko"

From: "Andrea Ring"

PERSON CARDINAL ORG LAW WORK_OF_ART GPE MONEY

> Top of Personal Folders > ring-a > Andrea_Ring_Jun2001 > Notes Folders > Sent

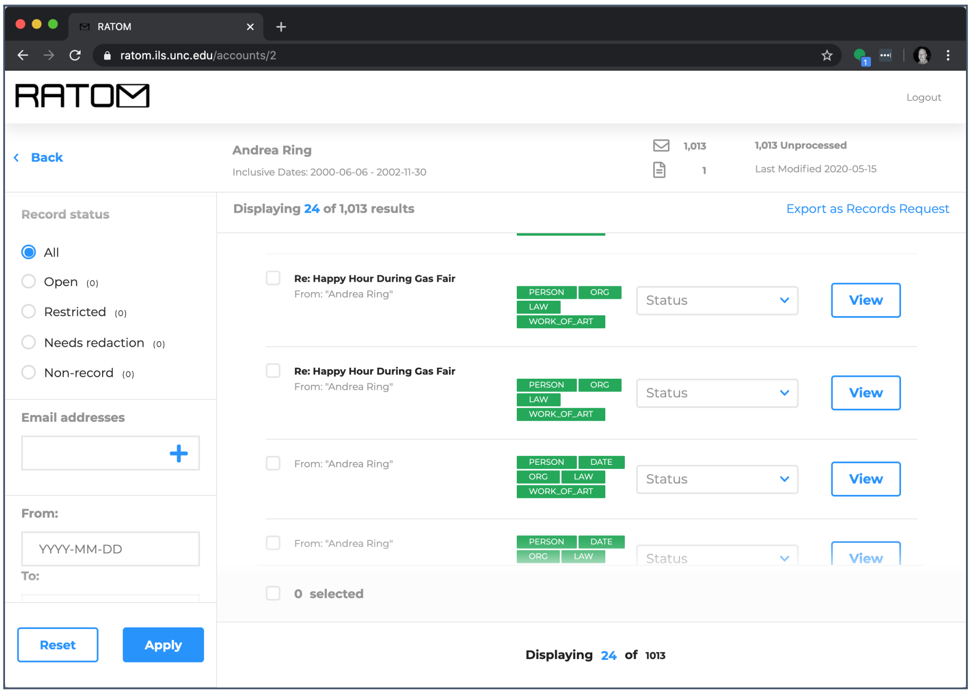
```
I do not show I did any deals outside of EOL with Koch at the Sabine Hub
during this time frame.  Sitara deal #369260 refers to EOL deal ID 369298.

***********
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc. This Data Set
is licensed under a Creative Commons Attribution 3.0 United States License <http: 3.0="" by=""
creativecommons.org="" licenses="" us=""></http:> . To provide attribution, please cite to "ZL Technologies, Inc.
(http://www.zlti.com)."
***********
```

37 of 1013

Figure 4. Message view in the RATOM web application.

The interface also provides Tagging and Search (Figure 5). These allow users to select by classification (e.g. record vs non-record) and date range. An audit history is maintained in the database to allow review and validation of tags applied at specific times or by specific users. Within a search result, the user may select all results in bulk or individual messages using the available checkboxes. Once the desired message set has been identified, an export report can be created using the “Export as Records Request” element near the top right. This generates a JSON report with elements corresponding to the files on the backing store and message IDs within those files. These can then be exported from the relevant PST(s) as individual EML files using the RATOM command-line tool, in preparation to serve a records request.



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

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

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

□ From: "Andrea Ring"

□ From: "Andrea Ring"

□ 0 selected

Displaying 24 of 1013

Figure 5. Tagging and search view in the RATOM web application.

The RATOM web application is intended to be deployed to cloud infrastructure, and is configured by default for deployment to Microsoft Azure. It is intended to be deployed to a managed Azure Kubernetes Service (AKS) cluster with at least two nodes and use an Azure database for PostgreSQL. It includes Ansible playbooks to automate deployment of both the server and the web application. Our public installation guide provides detailed instructions on how to configure the required services and accounts.

Sustainability and Reuse

The software developed for this project is open source (using the MIT license), distributed as source on GitHub, and in package form on the Python Package Index (PyPI). Our core library is cross-platform (Windows, Linux, and macOS), automatically built and tested in a CI workflow via GitHub Actions, and is automatically monitored for dependency updates and deprecations using Dependabot. Target constituencies should be able to find, use, and modify our tools as desired. This alone, however, is not enough to ensure sustainability.

The mechanisms for sustainability are the development and cultivation of a dedicated community of users who will deploy the software, share experiences and (when appropriate) contribute code revisions. In past TOMES and BitCurator efforts, we have observed a significant interest in the use of these tools by working librarians and archivists. The BitCurator Consortium (BCC) will serve as one of the mechanisms for sustaining tools produced as part of RATOM, along with facilitating growth and maintenance of the user community after the project is complete.

As a government institution, DAR has a statutory obligation to collect, preserve, and make accessible public records of the state of North Carolina, which includes the TOMES project. As such, DAR has a source of secure funding of support for our electronic records program and the long-term management of electronic records and projects. The TOMES project in particular will continue to be DAR's primary tool for the preservation, processing, and access to email as the institution continues to collect archival email accounts. There is an institutional commitment to maintain the software that is essential to our mandated services. In addition to ongoing maintenance and public access on GitHub, the tool and all associated code and documentation will be maintained on DAR servers, which are backed up and maintained at the North Carolina Department of Information Technology.

Outcomes

RATOM has strived to fulfil the principles laid out in the Santa Barbara Statement on Collections as Data (Padilla et al, 2018), to support the "computational use" of digital materials. We expect these resources to be beneficial both for GLAMs professionals (to carry out appraisal, description, reference services, discovery and review for sensitivity), and to end users who often care more about specific entities than what might exist within the bounds of a given storage medium or email account. We hope the RATOM project facilitates valuable and creative reuse of both email content and the software used to care for it.

Further development on the core libratom tools has continued at UNC SILS in 2022. All of our code is open source and available on GitHub[13], and released as packages on PyPI[14], along with associated build and use documentation. Interactive code notebooks intended to demonstrate functionality both to end users and developers are also available via our GitHub organization.

Conclusion

In this paper we describe a new toolset intended to facilitate email identification, analysis, and preservation. The project implemented

[13] https://github.com/libratom

[14] https://pypi.org/project/libratom/

software to scan email archive files (including PST, OST, and mbox) and record and export content, metadata, and derived features such as entities identified using natural language processing (NLP) in a SQLite database that can be queried as part of existing email processing workflows.

In addition to the core processing tool for email backup formats, the project developed a web-based iterative email processing tool to assist archivists in reviewing email collections for retention and release. This interface allows archivists to import email directly from PST, OST, and mbox formats, analyses message content to identify entities and other features of interest, and creates a set of processing accounts associated with each associated ingest.

Acknowledgements

The authors would like to acknowledge the individuals and organizations that have contributed to the design, development, testing, and funding of the RATOM project. RATOM was made possible by a grant from the Andrew W. Mellon foundation, supporting operations of the teams at UNC SILS and NC DAR between 2019 and 2020. Current and former staff members at NC DAR, including Camille Tyndall-Watson, Jamie Patrick-Burns, and Sangeeta Desai oversaw design and development of the RATOM web application. Eliscia Kinder and Anusha Suresh served as the project managers at UNC SILS. The Caktus Group in Raleigh, NC provided development services for the RATOM web application. Carl Wilson of the Open Preservation Foundation provided code review services. Finally, we would like to thank the members of the RATOM Advisor Board for their invaluable guidance and feedback: Euan Cochrane, Kevin DeVorsey, Glynn Edwards, Matthew Farrell, Lynda Schmitz Fuhrig, Kathleen Jordan, Joanne Kaczmarek, Kate Murray, Justin Simpson, Yue Wang, and Carl Wilson.

References

[book] Bromley, B. S., Christman, R. and Gray Eakin Page, S.. “I Really Can’t Wait to Archive This Exchange: Exploring Processing as Appraisal in the Time Kaine Email Project.” In Appraisal and Acquisition: Innovative Practices for Archives and Special Collections, ed. Kate Theimer. Lanham: Rowman and Littlefield, 2015.

[report] Chassanoff, A., Post, C., Skinner, K., Farrell, J., Locke, B., Perry, C., Smith, K., Wang, H., Lee, C. A., Meister, S., Meyerson, J., Rabkin, A., Zhang, Y., and Ballard, H., “OSSArcFlow Guide to Documenting Born-Digital Archival Workflows,” Educopia Institute, June 23, 2020, https://educopia.org/ossarcflow-guide/

[online magazine] Christman, R. & Gray Page, S., "Addressing the Challenge of the Governor's E-mail," MAC Newsletter 43, no. 1 (2010), https://lib.dr.iastate.edu/macnewsletter/vol43/iss1/10.

[journal article] Cocciolo, A. "Email as Cultural Heritage Resource: Appraisal Solutions from an Art Museum Context," Records Management Journal 26, no. 1 (2016): 68-82.

[thesis] Gilliland-Swetland, A., "Development of an Expert Assistant for Archival Appraisal of Electronic Communications: An Exploratory Study," PhD Dissertation, University of Michigan, 1995; Anne J. Gilliland, "Designing Expert Systems for Archival Evaluation and Processing of Computer-Mediated Communications." In Research in the Archival Multiverse, edited by Anne J. Gilliland, Sue McKemmish and Andrew J. Lau, 686-722. Clayton, Australia: Monash University Publishing, 2016.

[report] Harvey, R. & Thompson, D., "Automating the Appraisal of Digital Materials," Library Hi Tech 28, no. 2 (2010): 313-22.

[report] Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., and Varner, S., “The Santa Barbara Statement on Collections as Data,” May 2018, https://doi.org/10.5281/zenodo.3066209.

[report] Prom, C. J. “Preserving Email - DPC Technology Watch Report 11-01.” Digital Preservation Coalition, December 1, 2011. http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf.

[journal article] Smith, D., ”Thinking Outside the Box: Use of Predictive Coding as a RIM Tool,” Information Management 47, no. 1 (2013): 30-32,46.

[report] Task Force on Technical Approaches for Email Archives. "The Future of Email Archives." Washington, DC: Council on Library and Information Resources, 2018.

[report] "The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond." National Archives of the UK, 2016.

[journal article] Vinh-Doyle, W. P. “Appraising Email (Using Digital Forensics): Techniques and Challenges," Archives and Manuscripts 45, no. 1 (2017): 18-30, DOI: 10.1080/01576895.2016.1270838

[journal article] Wang Y., Zheng, K., Xu, H., and Mei, Q., “Interactive Medical Word Sense Disambiguation through Informated Learning,” Journal of the American Medical Informatics Association 25, no. 7 (2018): 800-808.