Using ePADD to Process Email Archives

EA:BCC Email Symposium
June 13, 2023

Sally DeBauche

Sally DeBauche is a Digital Archivist in the Department of Special Collections at Stanford University Libraries. She is responsible for creating policy and workflows related to born digital archiving and processing born digital collections, with a particular focus on email. She also project managed the development for the ePADD software from 2020-2021 and consulted on the most recent cycle of development led by the University of Manchester and Harvard University. Sally received a BA in History from the University of Wisconsin-Madison and an MSIS from the University of Texas at Austin.

Workshop outline

- Introducing ePADD
- Hands on with ePADD - Basics

Break

- Hands on with ePADD - Advanced
- How to participate in the ePADD community
- Additional resources

Why ePADD?

ePADD helps cultural memory institutions meet the challenges of collecting, processing, and providing access to email

- Screening email for sensitive, restricted, or legally protected information
- Providing browsing access to intellectual content of collection
- Enabling research and scholarship

Free & open source software

- Incorporates machine learning, automated metadata extraction, and natural language processing
- Licensed under Apache 2.0

Project partners

Stephen Abrams, Head of Digital Preservation, Harvard University

Tricia Patterson, Senior Digital Preservation Specialist, Harvard University

Ian Gifford, Library Application Development Manager, University of Manchester

Jochen Farwer, Application Developer, University of Manchester

Tom Higgins, Application Developer, University of Manchester

Jessica Smith, Creative Arts Archivist, University of Manchester

Paul Carlyle, Curator for Nonconformist Collections, University of Manchester

Glynn Edwards, Assistant Director, Department of Special Collections, Stanford University

Michael Olson, Service Manager, Born-Digital Preservation Lab, Stanford University

Sally DeBauche, Digital Archivist, Stanford University

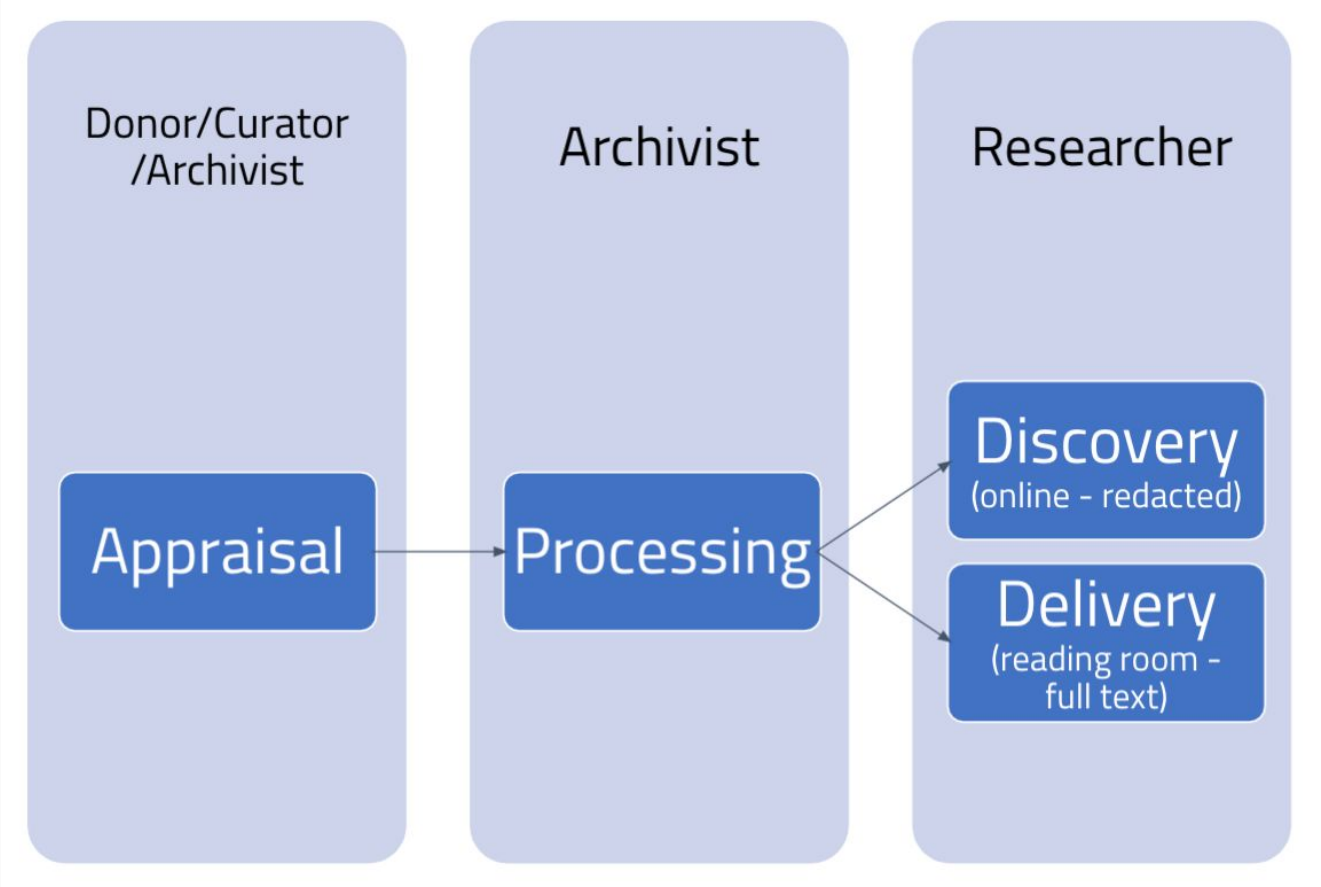
Project funders

User community

British Library
Brown University
California Inst. of Technology
Canadian Centre for Architecture
Center for Jewish History
Columbia University
Duke University
Emory University
Fordham University
Getty Research Institute
Harry Ransom Center
Harvard University
Indiana University - PUI
Mass. Inst. of Technology
Museum of Modern Art
National Library of New Zealand
New York Philharmonic
New York University
Princeton University
Rockefeller Archive Center
Royal Library of Copenhagen
Smith College
Stanford University
Tufts University
University of California, Berkeley
University of California, Irvine
University of California, LA
University of California, Santa Cruz
University of Copenhagen
University of Illinois – UC
University of Manchester
University of Minnesota
University of Southern California
University of Texas at Austin
University of Virginia
University of Warwick
Wildlife Conservation Society
92nd St. YMCA, NYC

ePADD

ePADD architecture - the four modules



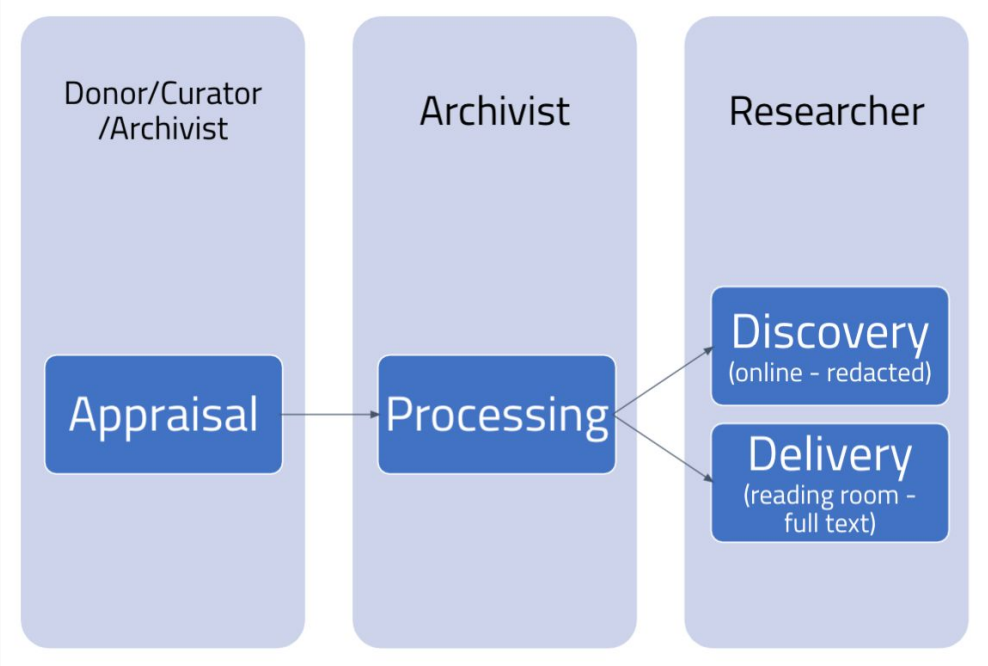
ePADD architecture - the four modules

Appraisal
- Supports email ingest and optional donor/curator review

Processing
- Supports review and description by staff of cultural heritage organizations
- Supports association of email correspondents with authorized name headings

Discovery
- Supports public access to redacted email (only includes named entities, message dates, and name part of email addresses)

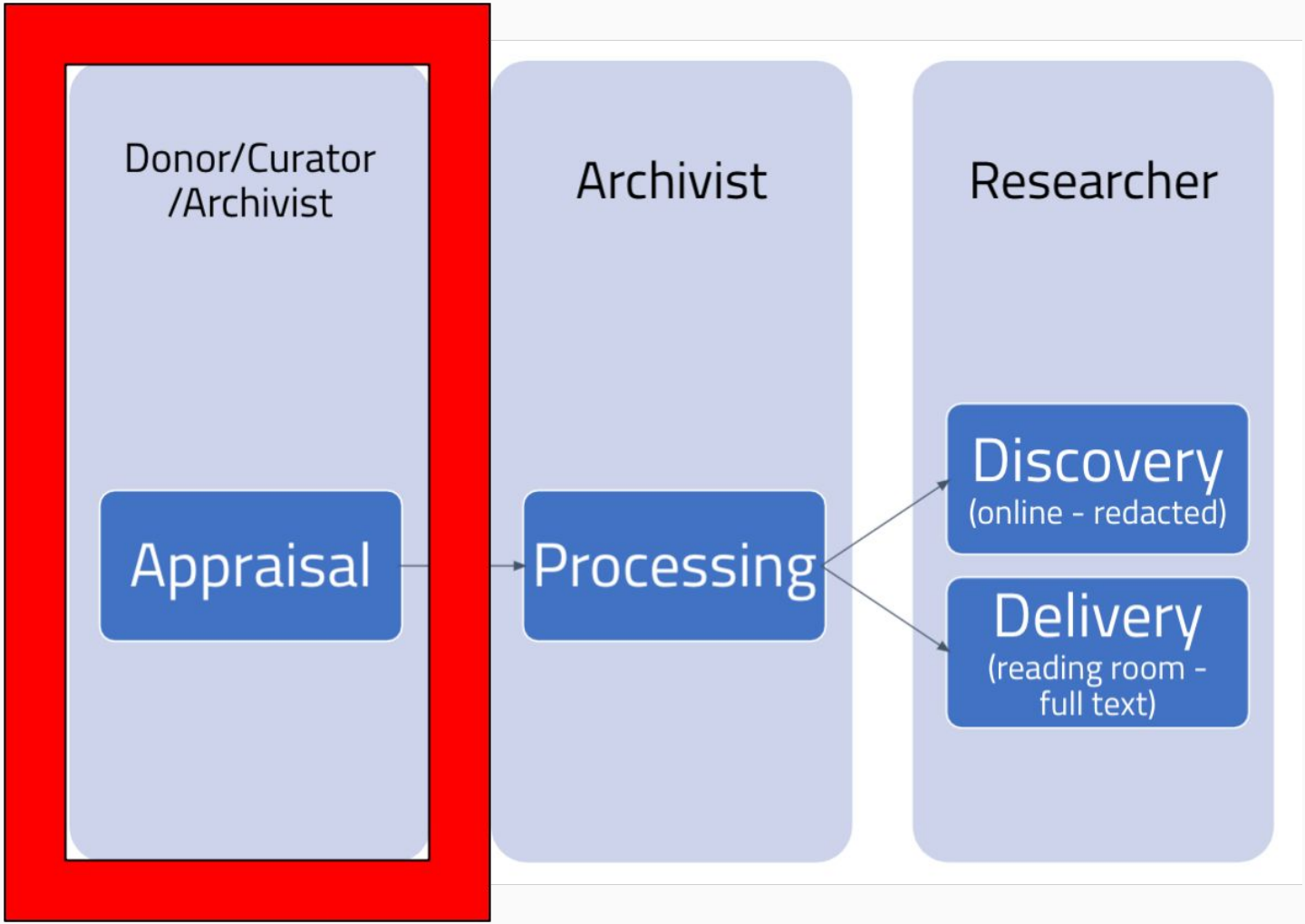
Delivery
- Supports researcher access to all (non-restricted) messages, including attachments



ePADD architecture - note

- The functionality of each module is geared toward the target user and workflow stage
- Not all functionality is the included in every module
- We will be exploring the software from the perspective of all users but most users will only use the Appraisal and Processing Modules

Hands on with ePADD


Donor/Curator
/Archivist
Archivist
Researcher
Appraisal
Processing
Discovery
(online - redacted)
Delivery
(reading room -
full text)

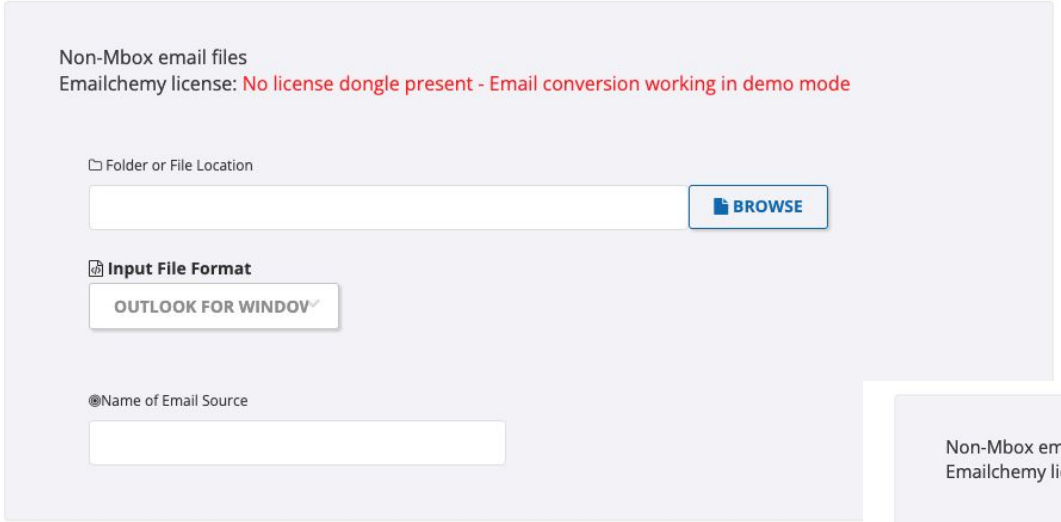
Email ingest

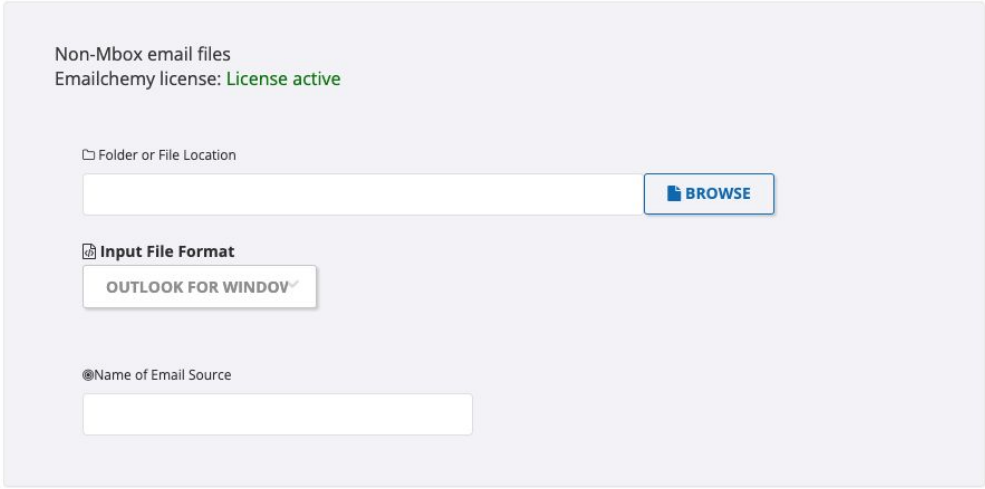
- ePADD accepts collections as .mbox file(s), other file types (with Emailchemy license) or via IMAP connection
- Multi-factor authentication can complicate ingest of email via IMAP connection

Emailchemy for ePADD

- Emailchemy is an email format conversion tool – can handle most common file formats
- ePADD v10 includes optional Emailchemy integration
- Emailchemy for ePADD license
  - $150 for first year, $99 each year thereafter
  - Uses dongle installation

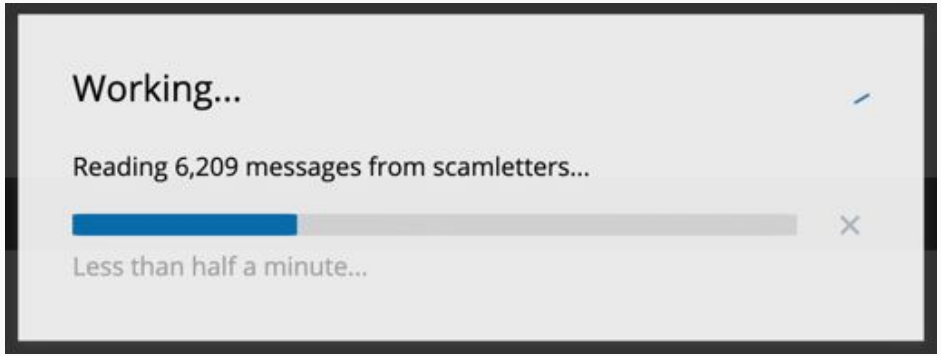
Emailchemy for ePADD





Email ingest process

- During ingest, ePADD takes the following actions:
  - Message deduplication
  - Entity recognition
  - Correspondent name resolution
  - Converting email files (optional)

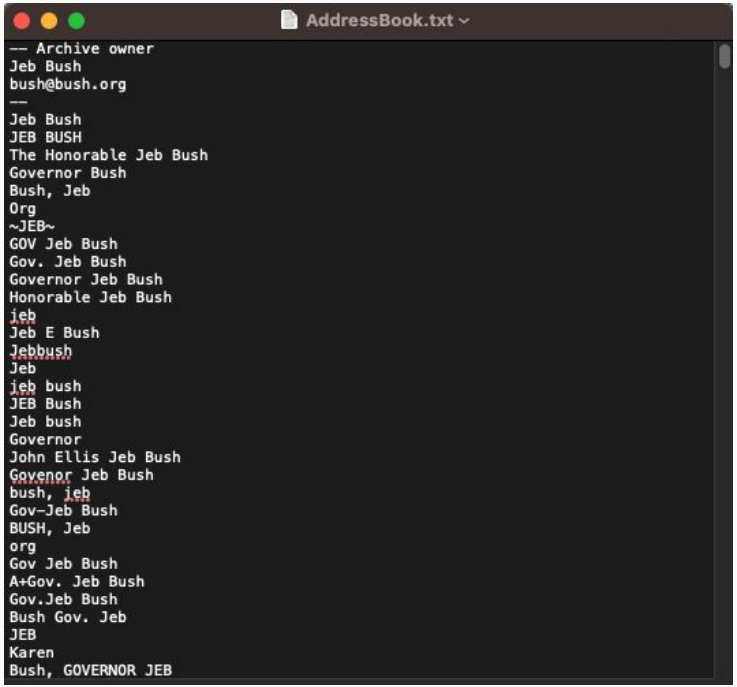


Entity recognition

- ePADD recognizes fine-grained entities for some entity types bootstrapped from DBpedia
- Examples of fine-grained entity types extracted by ePADD include:
  - Persons
  - Government Bodies
  - Events
  - Awards
  - Diseases

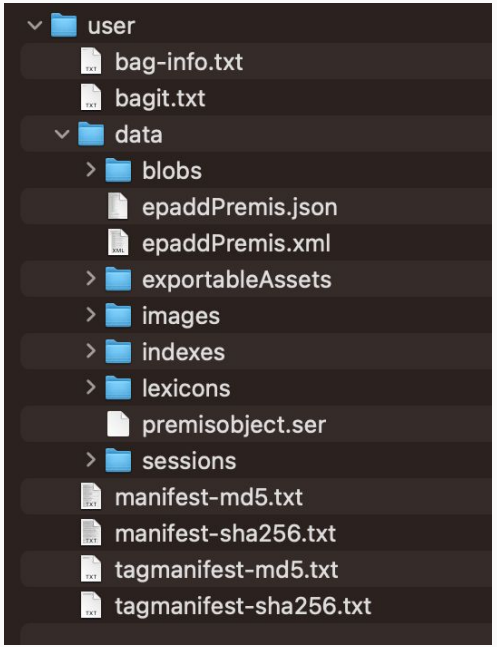
Name resolution

- ePADD concatenates names and email addresses associated with a single individual
  - Correspondent names with the same email address are grouped together
  - Correspondent names which are algorithmically identified as sufficiently similar are grouped together
- Users can edit the results of name resolution



ePADD data storage

- ePADD stores all collection data in ”user” directory
  - On Windows: C:\Users\<username>\
  - On Mac: User/<username>/



Indexing reports

ePADD provides information on the date and time of an import, the names of email sources, and a summary of any error types

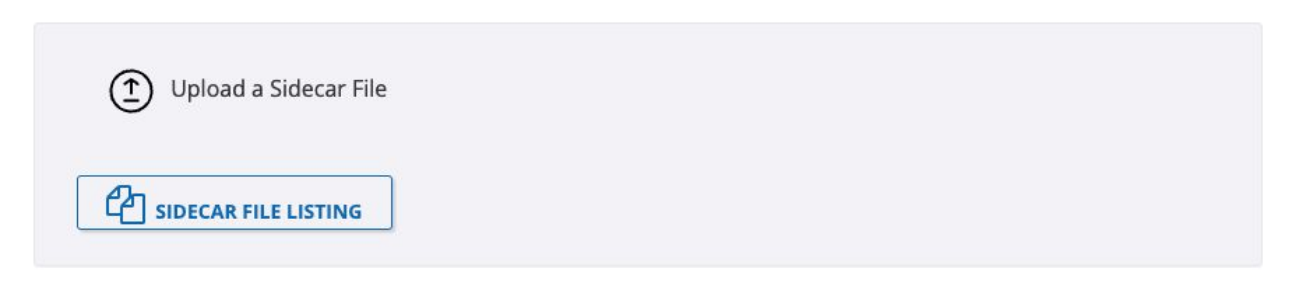
Jump to Errors

Import #1

Import Date: 2023-03-08 12:53:44 Pacific Standard Time
Fetch and index time: 49 seconds
Archive Owner: Jeb Bush
Primary Email Address: bush@bush.org
Archive Title: Jeb Bush Test Archive
Email Sources:
Selected folders: 3
: (mbox:/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/12 December 2003 Public 2.pst/12 December 2003/Top of Outlook data file.mbox messages:753)
: (mbox:/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/07 July 2003 Public 2.pst/07 July 2003/Top of Outlook data file.mbox messages:1518)
: (mbox:/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/10 October 2002 Public.pst/October 2002 Public -2-/Top of Personal Folders.mbox messages:1855)
selected_messages: 0, filtered: 0, imported: 4,124, duplicates: 2
No message filter
Messages span: 2000-07-31 to 2003-12-31
Space saved from detected duplicate messages: 1KB
Space saved from detecting duplicate attachments: 8,685KB

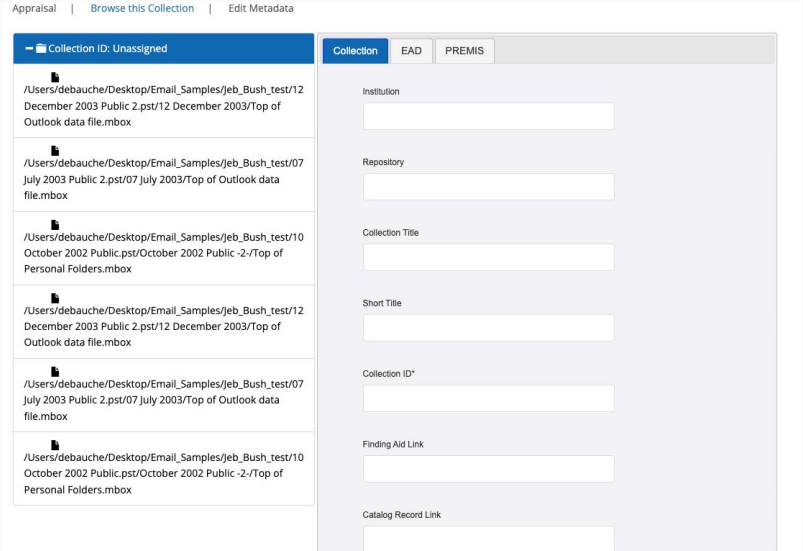
Ingesting sidecar files

- Sidecar files can be added to your collection when you ingest it into ePADD
- Sidecar files can be of any format
- Sidecar files support your collection management and documentation



Reviewing and adding metadata

- Users can review and add the following metadata:
  - Collection metadata to display in Discovery & Delivery Modules
  - EAD metadata
    - Traditional finding aid data
    - Processing notes
    - Embargo dates
  - PREMIS XML metadata
    - Events during ingest
    - Appraisal events

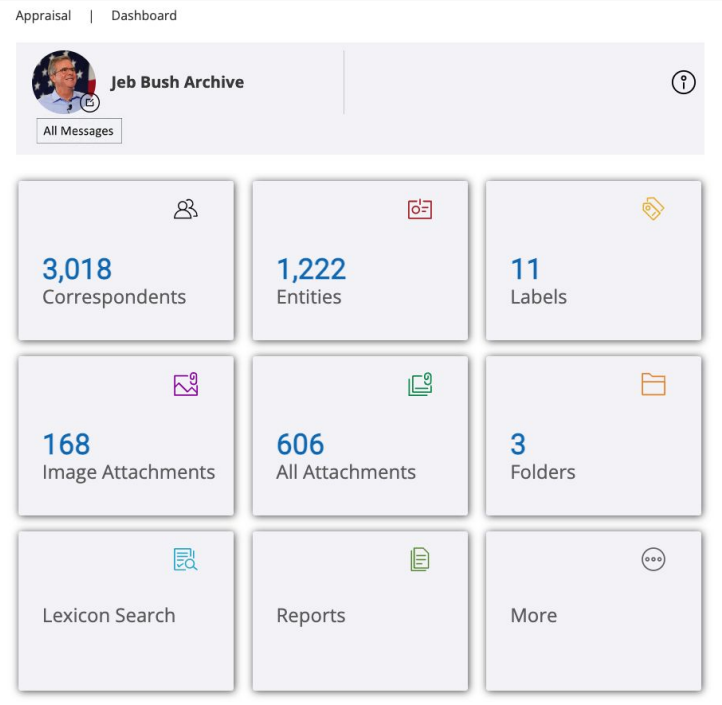


Adding PREMIS events

- Users can add PREMIS events
  - More Screen > “PREMIS Metadata”
- Events stored in epaddPremis.xml and epaddPremis.json files

Browse functions

- Users can browse a collection by:
  - Correspondents
  - Entities
  - Labels
  - Image attachments
  - All attachments
  - Folders
  - Lexicons



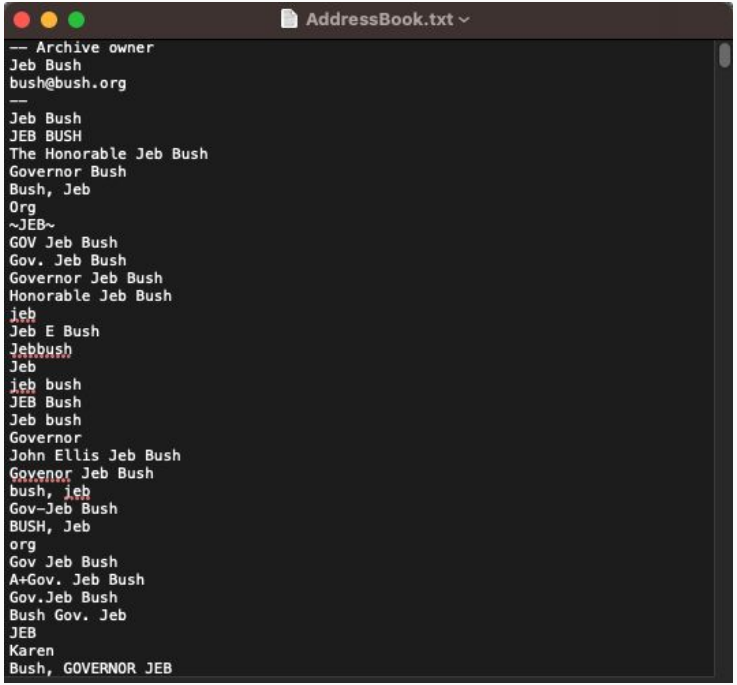
Browsing correspondents

- Enables users to browse messages associated with correspondents alphabetically or by volume (sent/received/received from donor)
- The address book can be edited by the users to fix names that ePADD has resolved incorrectly, or to identify addresses with mailing lists
- The address book can be debugged to investigate problems
- An account owner can optionally be set (will reflect in messages “received from owner” statistics)

Exercise: edit the address book

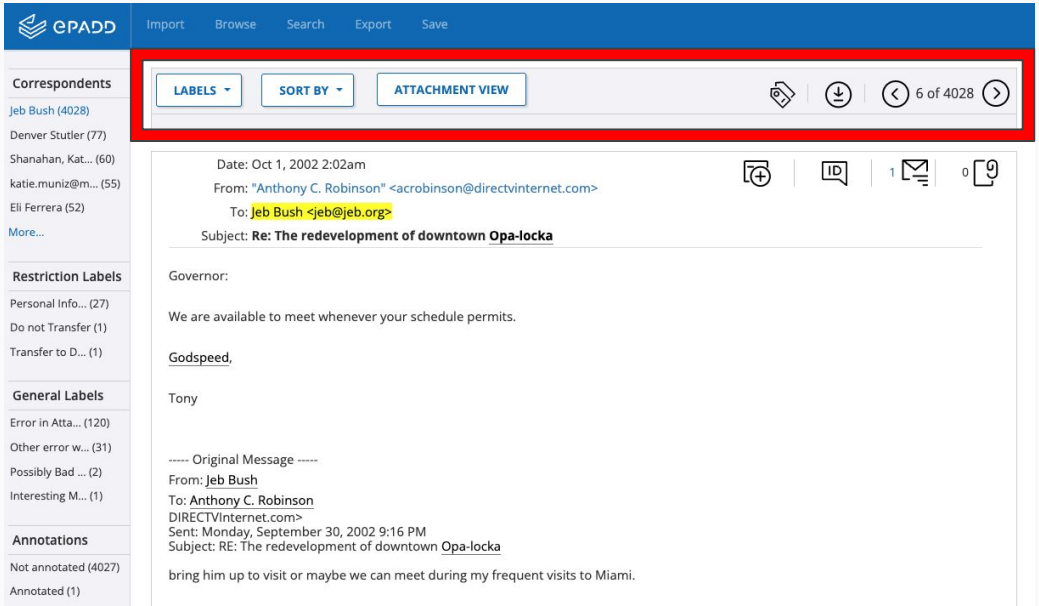
1. Go to Browse>Correspondents
2. Download address book
3. Edit address book
4. Upload edited address book

★ Question: Why would you edit the address book and how would you approach this process?



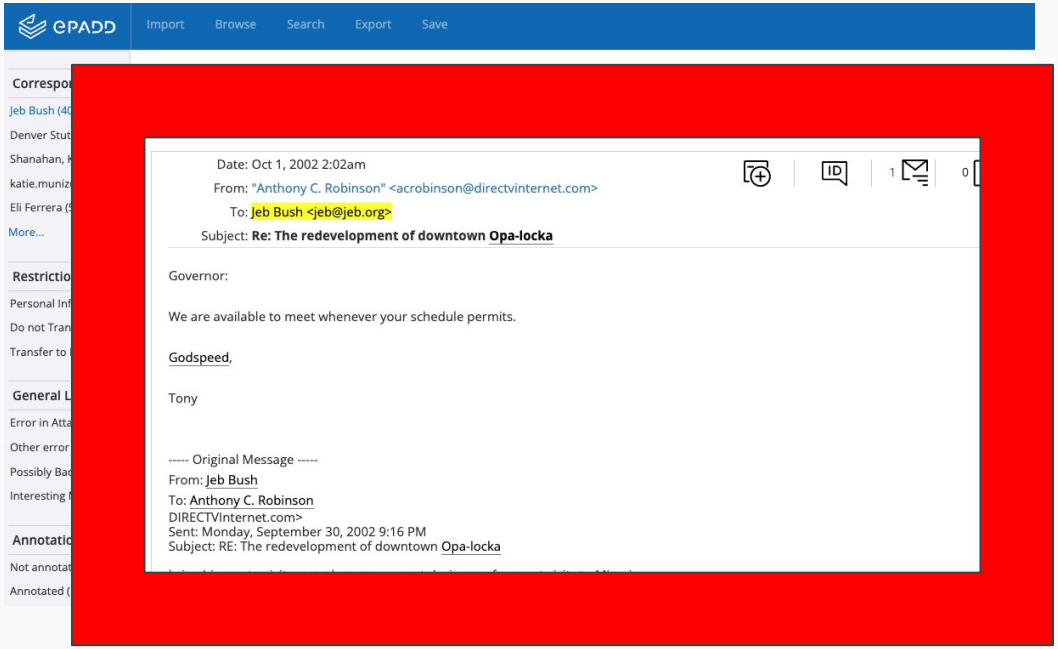
Message view

- Add label(s)
- Change sort order
- Switch to Attachment View
- Bulk labeling
- Download messages (.mbox format)
- Forward/Backward



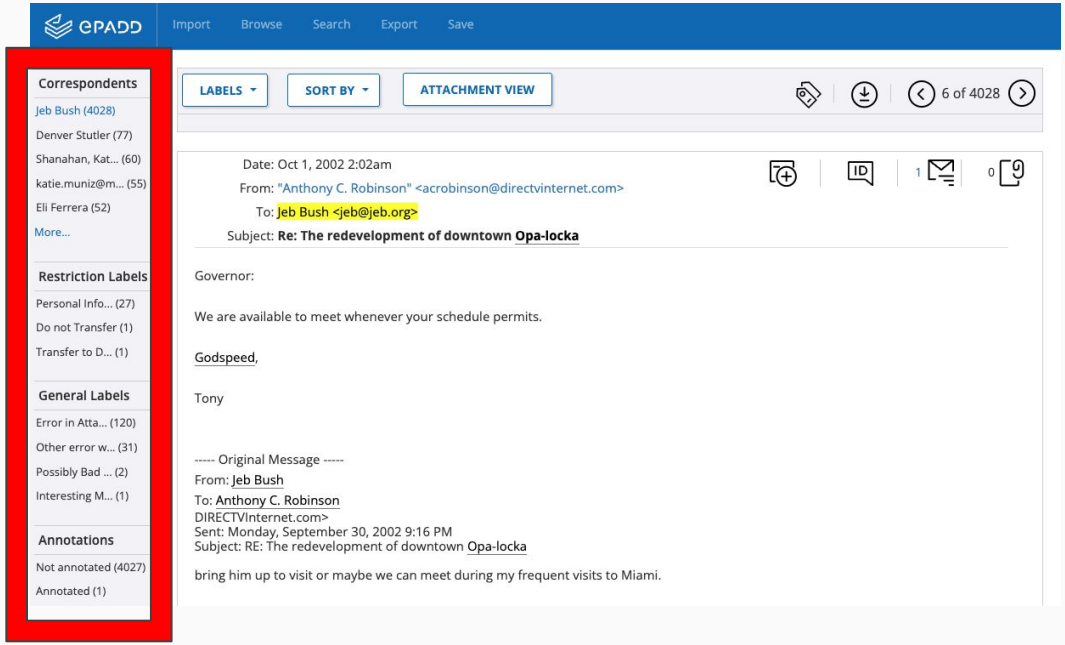
Message view

- Email header
- Add message annotation
- Message ID
- Open email thread
- Jump to attachments
- Email body & attachment(s)



Message view

- Facet messages
- Save!



Browsing entities

- ePADD uses some of the entities created in Wikipedia as the training data set for fine-grain entity extraction
- When the trained model is applied to messages, there will be results which have entries in Wikipedia (showing a score of 1)
- Predicted entities that do not have entries in Wikipedia have a score of <1

Labelling messages

- Users can assign customized labels to messages or sets of messages based on various criteria
- Labels are classified as:
  - System labels (used to support error identification)
  - General Labels
  - Restriction labels (which may be actionable or non-actionable)
  - Transfer to Delivery Only label

Exercise: creating and applying labels

1. Go to Browse>Labels
2. Select “Create Label”
3. Create 1 “General Label” and 1 “Restriction Label”
4. Go to Browse>“All Messages”
5. Apply a General Label to 1 message
6. Apply a “Restriction Label” to 1 message
7. Apply a “Do Not Transfer Label” to 1 message
8. Save!!!
9. Note the Message ID for each message in a text document outside ePADD – we will revisit this later in the Processing Module

Browsing image attachments

- ePADD includes an in-application viewer for image attachments
- Users can facet attachments by file type, file size, associated correspondent, and associated folder
- Users can view attachments in a grid view with thumbnails or in a list view
- Selecting an image brings the user to all messages to which that image is attached

Browsing all attachments (images & all other files types)

- Users can browse all attachments chronologically
- Some file types can be viewed within ePADD – for other attachment types, users can download and use third party software (e.g. QuickView Plus) to render them
- Users can view attachments in a grid view with thumbnails or in a list view
- Selecting an attachment brings the user to all messages to which that file is attached

Browsing folders

- Users can browse messages associated with a certain source file or folder

Source	Folder	Messages from Owner	Total
mbox	/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/10 October 2002 Public.pst/October 2002 Public -2-/Top of Personal Folders.mbox	0	1855
mbox	/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/07 July 2003 Public 2.pst/07 July 2003/Top of Outlook data file.mbox	0	1516
mbox	/Users/debauche/Desktop/Email_Samples/Jeb_Bush_test/12 December 2003 Public 2.pst/12 December 2003/Top of Outlook data file.mbox	0	753

Showing 1 to 3 of 3 entries

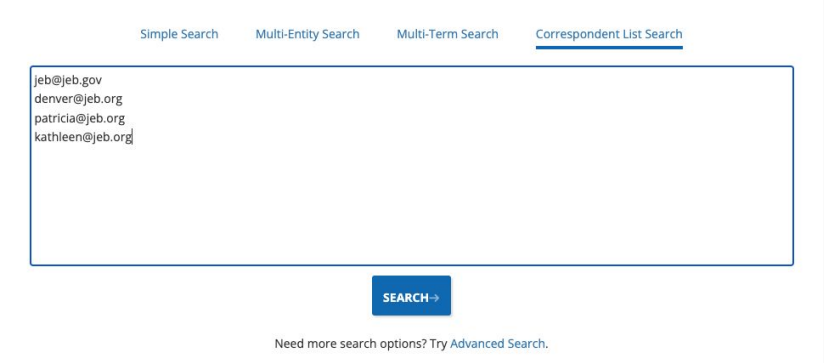
Previous Next

Search functionalities

- Keyword search
- Multi-entity search
- Multi-term search
- Correspondent search list
- Lexicons (including regular expression search)
- Advanced search

Correspondent list search

- Use Case 1: Donor gives you a list of email address and asks you to restrict all associated messages
- Use Case 2: Researchers want to find messages associated with a list of email addresses
- Email addresses should be entered (or pasted) one per line

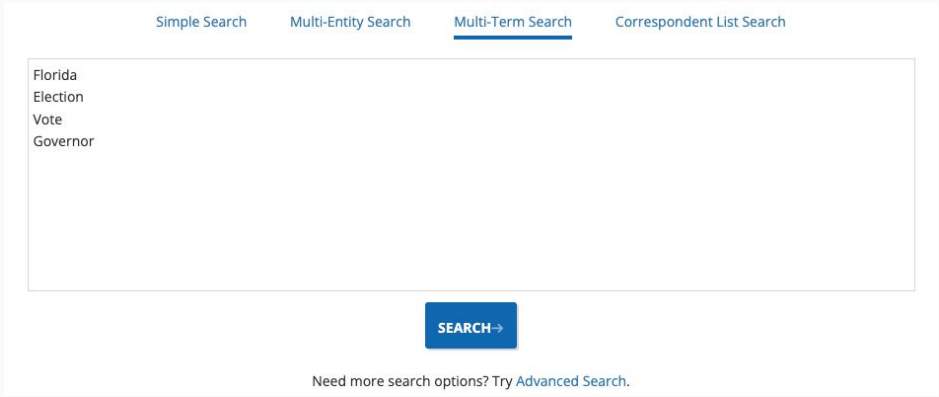


Multi-entity search

- Type or copy/paste in a block of text
- ePADD will identify terms that are present in the collection as extracted entities

Multi-term search

- Search on multiple terms at the same time
- Enter one term per line
- Efficient way of search lots of terms or similar terms

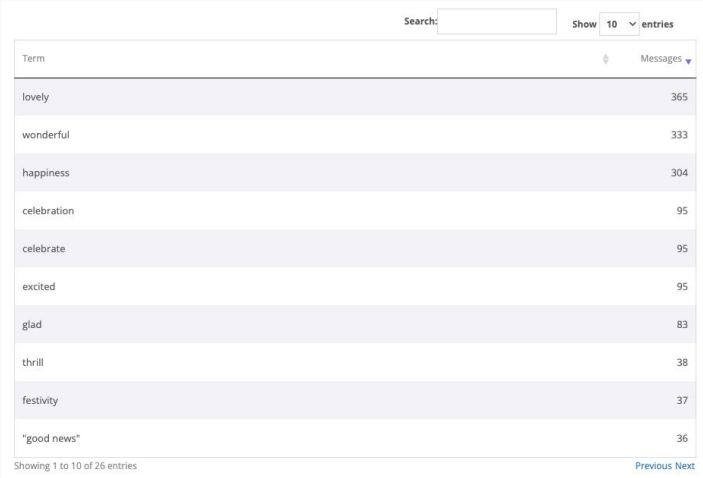


Lexicon search

- Lexicons are groups of keywords, organized by commonly searched categories that can be searched in bulk
- ePADD includes several pre-loaded lexicons, including lexicons geared toward searching for sensitive information and regular expressions
- All lexicons are customizable
- Users can also add their own lexicons

Testing a lexicon category

- Users can test the appropriateness of the terms in a lexicon category by displaying the number of search hits for each term



Regular expression search

- Identifies data that is in a consistent pattern such as credit card numbers and U.S. Social Security Numbers
- In ePADD, each pattern is its own lexicon category
- Like other lexicons in ePADD, this one is customizable

How to screen for sensitive information

- Sensitive Lexicon
- Regular Expression Search
- Disease Entity Browsing
- Image Attachment Browsing
- Correspondent List Search
- Advanced Search to combine tactics!

Exercise: create a new lexicon

1. Go to Browse>Lexicon Search
2. Select “Create Lexicon”
3. Add terms with “|” separating each term, using quote marks to encapsulate multi-word terms

★ Tip: Try using the “test category” function to refine your terms

Restricting materials using labels

- Messages can be restricted until a specific date
- Messages can be restricted for a specific length of time from the date of sending of the message
- Messages can be restricted until further notice (i.e. for an undefined period)
- Messages can be assigned more than one restriction label
- Messages can be restricted to different degrees
    - “Do Not Transfer” labelled messages will not export from the Appraisal Module
    - “Transfer to Delivery Only” labelled messages will export to the Delivery Module but not the Discovery Module

Releasing restricted messages

- Messages with a restriction label are only released to the next module if both of the following are true:
  - The restriction period for the message has expired
  - The “Cleared for Release” label has been assigned to the message
- After releasing previously restricted messages, the archive will need to be re-exported to the Discovery and Delivery Modules and the respective folders will need to be updated

Export options

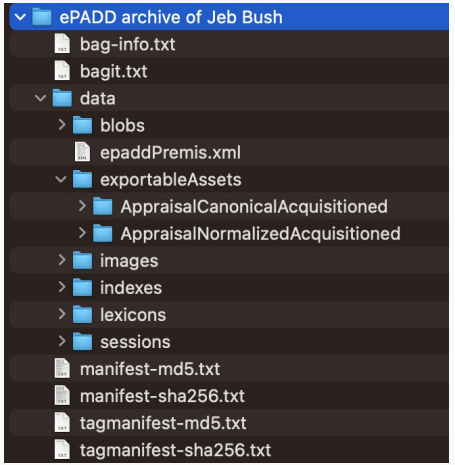
- Users can export:
  - A preservation ready package
  - To the next ePADD module
  - As an .mbox file
  - The original text (supports topic modeling)
  - Attachments not recognized by Apache Tika
  - A list of correspondents
  - Message headers (supports network analysis)
  - Entities

Exporting a preservation package

- Can optionally include:
  - Appraised/processed .mbox or .eml files
  - Unprocessed canonical email file(s)
  - Metadata
  - Additional “sidecar” file(s)

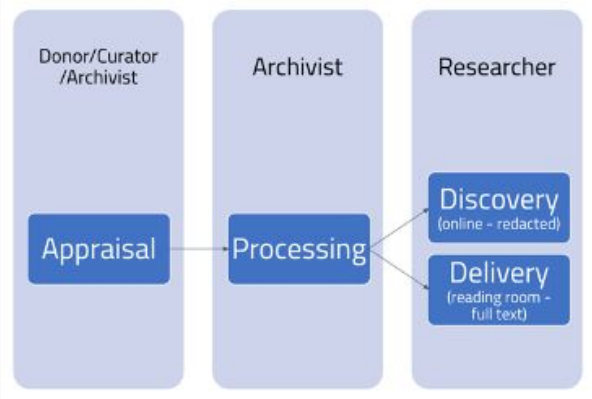
Export to preservation

- Data is saved here:
  - On Windows:
    C:\Users\<username>\epadd-appraisal\user\data\exportableAssets\
  - On Mac:
    User/<username>/epadd-appraisal/user/data/exportableAssets/
- Definitions:
  - AppraisalCanonicalAcquisitioned
    - Files before conversion with Emailchemy
    - Untouched imported .mbox files
  - AppraisalNormalizedAcquisitioned
    - Mbox files generated by Emailchemy
    - Untouched imported .mbox files
  - AppraisalNormalizedAppraised
    - Appraised emails in their respective folders
    - Can be in .mbox or .eml format
    - Includes labels and annotations



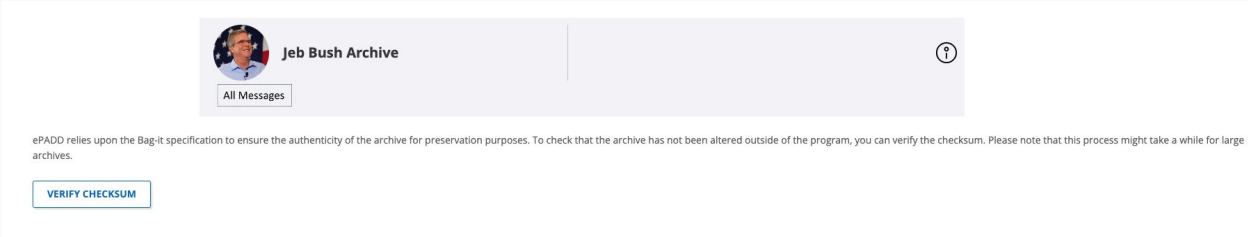
Exporting to next ePADD module

- Appraisal Module Processing Module
- Processing Module Discovery & Delivery Modules (2 versions of processed collection)
- ePADD uses BagIt to ensure file integrity between modules



BagIt file packaging format

- The format used by the Library of Congress and many digital libraries to verify files during transfer
- Users should use the “Verify Checksum” function to verify the integrity of the content transferred between modules



Export original text of messages

- ePADD also supports exporting the “original text of all non-restricted messages as .txt files” for text analysis, including topic modeling
- The “original text” of an email message is the text entered by the creator of that message, not including message replies or forwards

Export messages as .mbox or .eml

- Enables users to export all messages, all unrestricted, or all restricted messages as an .mbox or .eml file
- Use Case 1: Researchers using the Delivery Module in a reading room use this function to download messages an an .mbox or .eml for their research and coordinate with Public Services for access to the download
- Use Case 2: Public Services staff using the Delivery Module in a reading room use this function to download messages a an .mbox or .eml to meet a request from a remote researchers
- Use Case 3: Archivist downloads collections as .eml file as desired format for preservation

Export attachments not recognized by Apache Tika

- ePADD uses Apache Tika to detect and extract metadata and text for over one thousand different file types
- Extracted text is indexed by ePADD for full text search
- If Apache Tika cannot recognize a file type, that file will not be included in full text search
- Users may choose to export these unrecognized attachments to review them with another tool

Use Case: Export the unrecognized attachments and review them with Bulk Extractor or Identity Finder. You find some additional attachments that contain PII and apply restriction labels in ePADD

Export correspondents

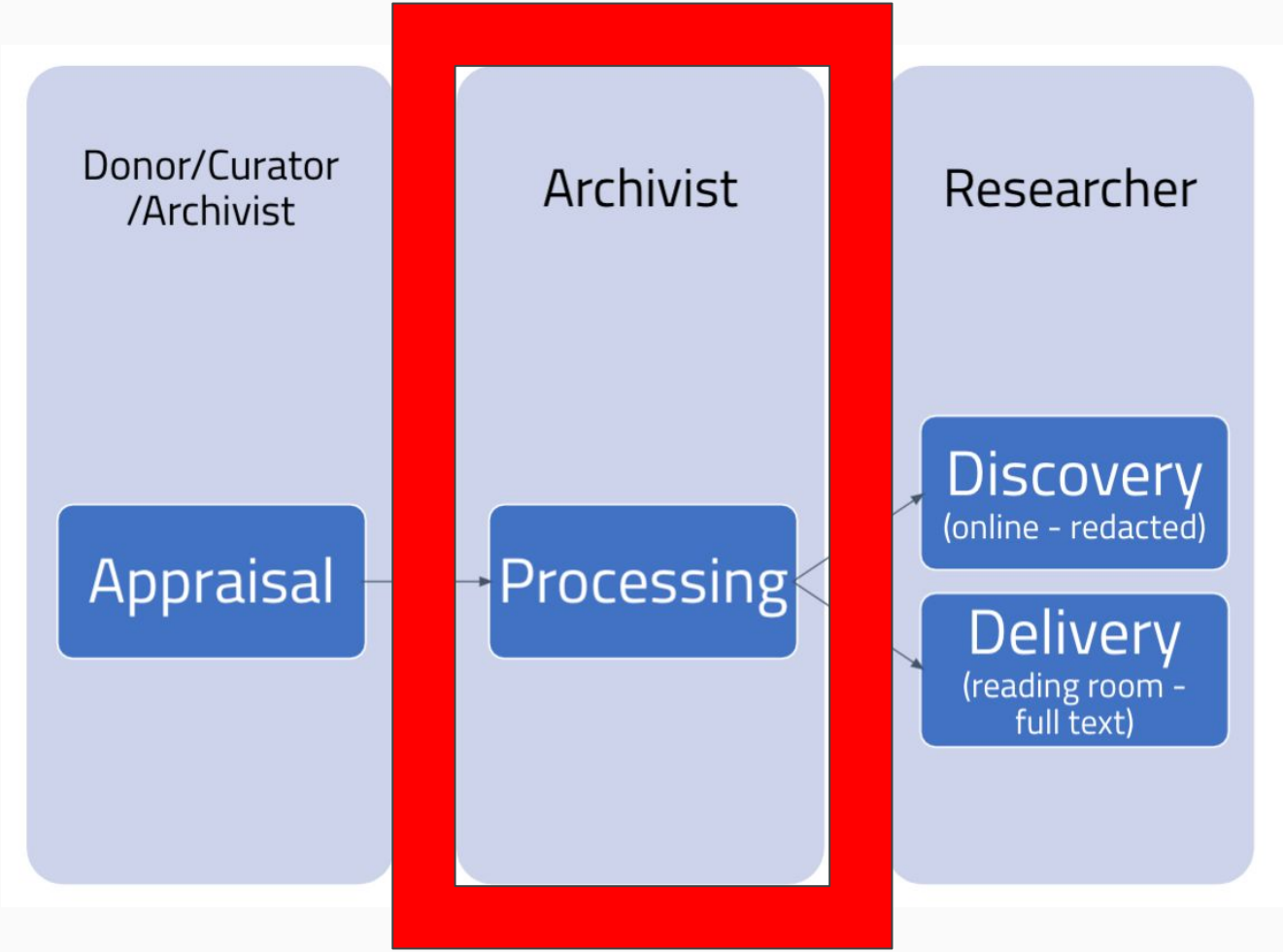
- Users can export:
  - All correspondents
  - Only correspondents that have been associated with an authority heading in the Processing Module
  - Only correspondents that have not been associated with an authority heading in the Processing Module

Export message headers

- Exported message headers include:
  - Message date
  - To email / name
  - From email / name
  - CC email / name
  - BCC email / name

Export entities

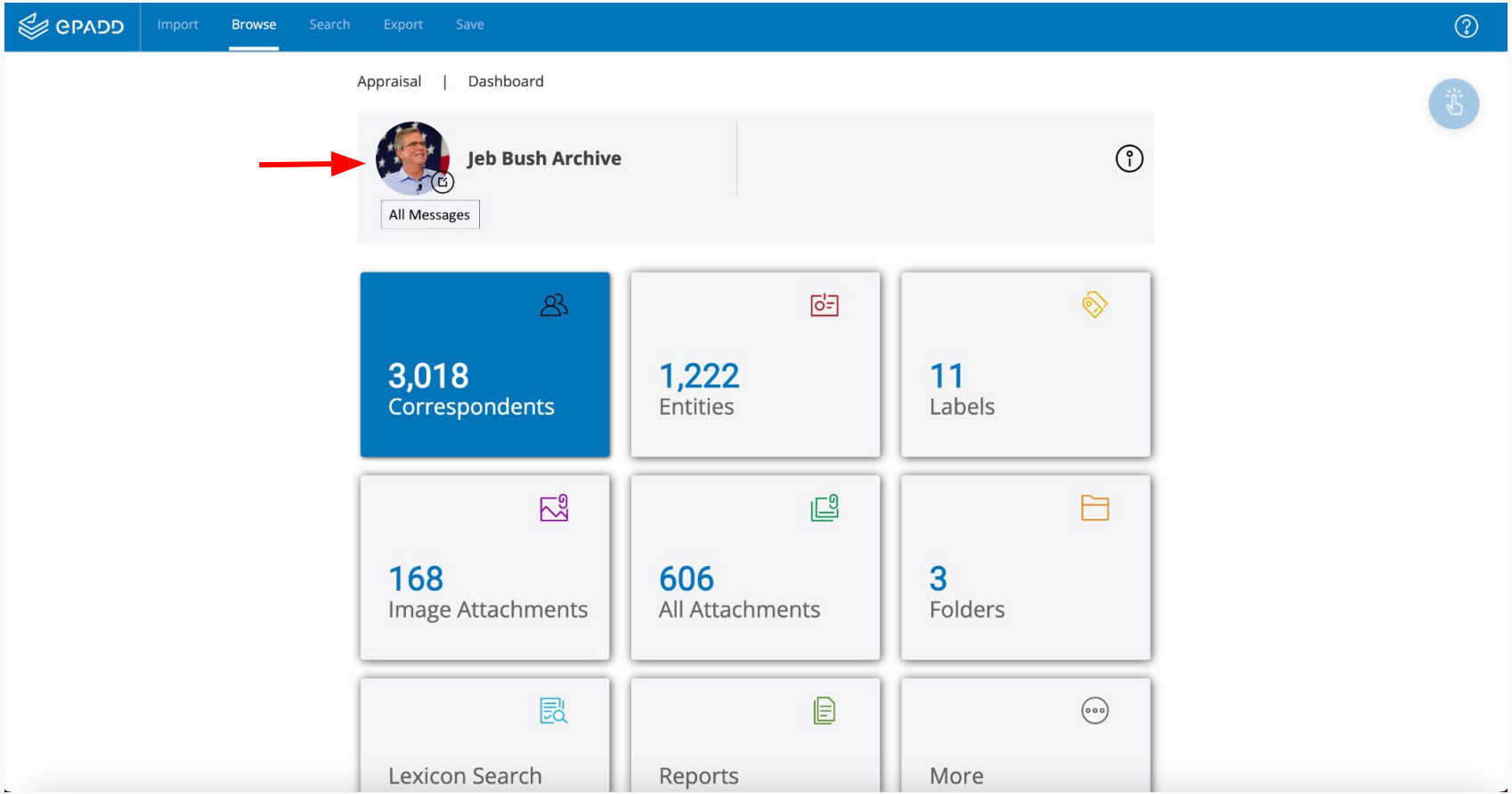
- ePADD supports exporting:
  - All entities
  - Only entities from a particular category


Donor/Curator /Archivist
Archivist
Researcher
Appraisal
Processing
Discovery
(online - redacted)
Delivery
(reading room - full text)

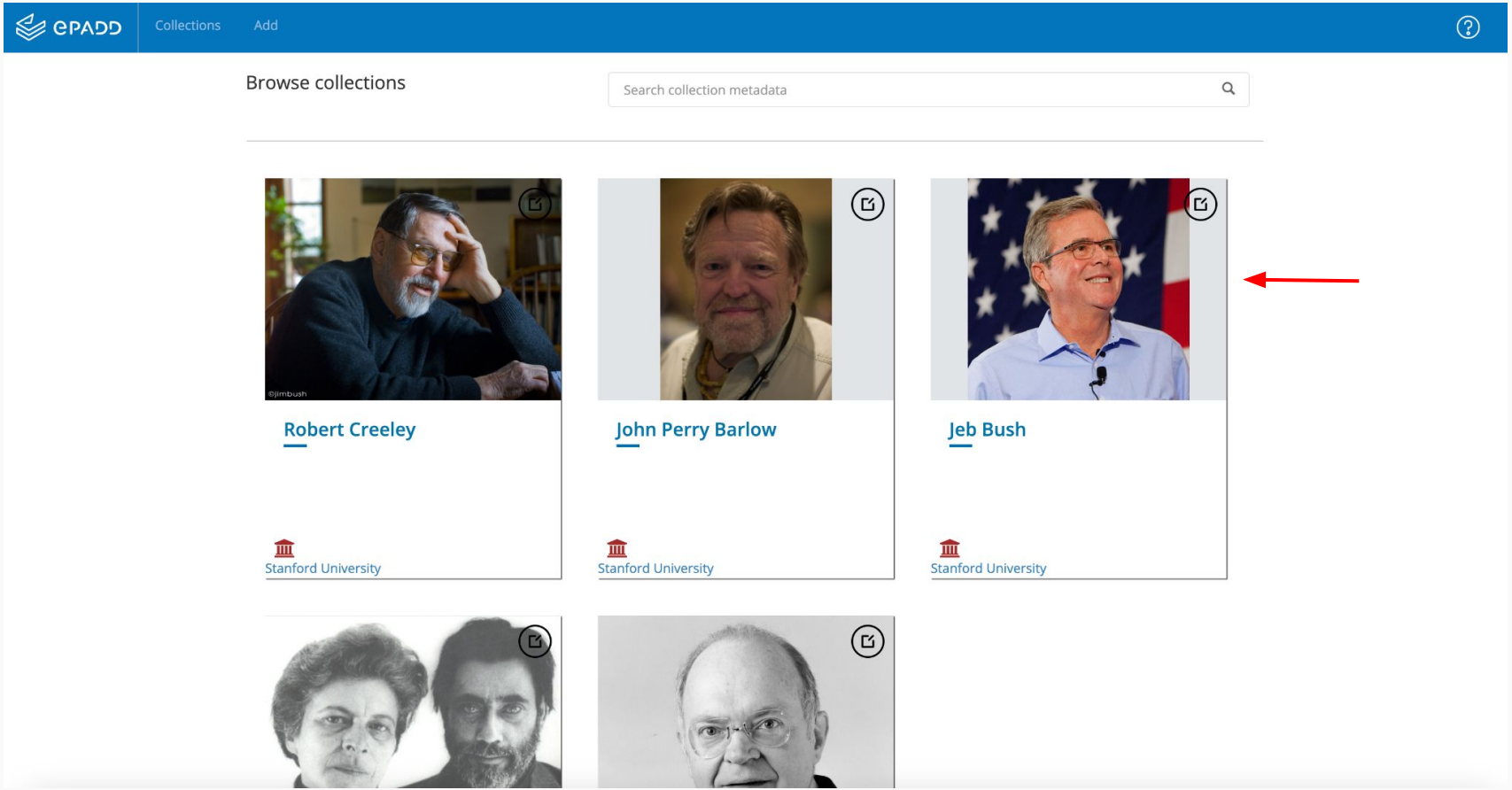
Adding collection images to enhance user experience

- Profile Image – The image shown at the top left-hand corner of many pages in the Appraisal, Processing, Discovery, and Delivery Modules
- Landing Image – The image shown in the “Collections” page in the Processing, Discovery, and Delivery Modules
- Banner Image – The image shown under the “About this Collection” in the Processing, Discovery, and Delivery Modules

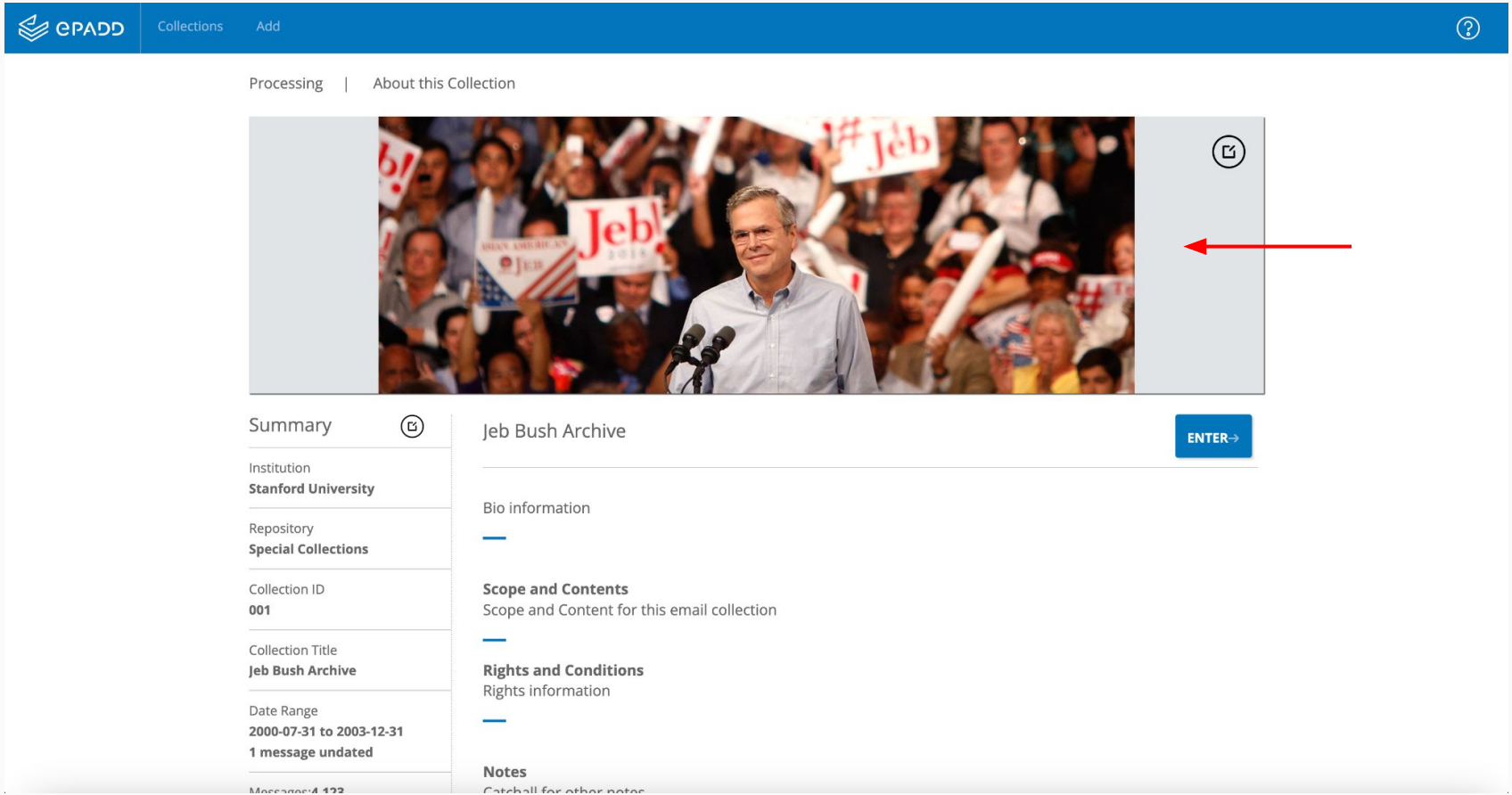
Profile image



Landing image

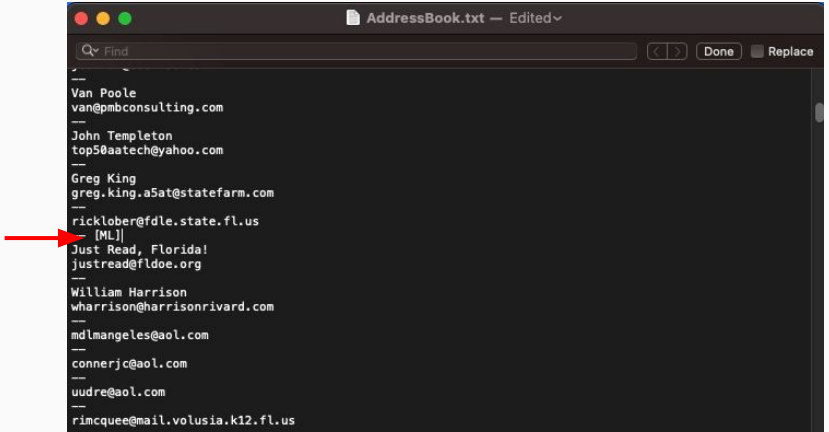


Banner image



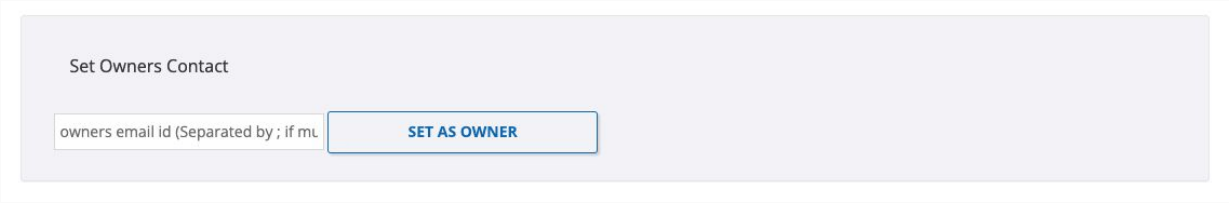
Identifying mailing list addresses

- ePADD will guess which addresses belong to mailing lists
- [ML] will be assigned to these addresses
- Users can manually assign [ML] to correspondents in the addressbook
- Users can include “mailing list” as one of the filters to perform advanced searches



Setting the owner’s contact

- Setting the primary email address is not mandatory
- Setting the owner’s email is used to calculate the number of messages received from the designated owner
- This is especially useful for archives with multiple owners

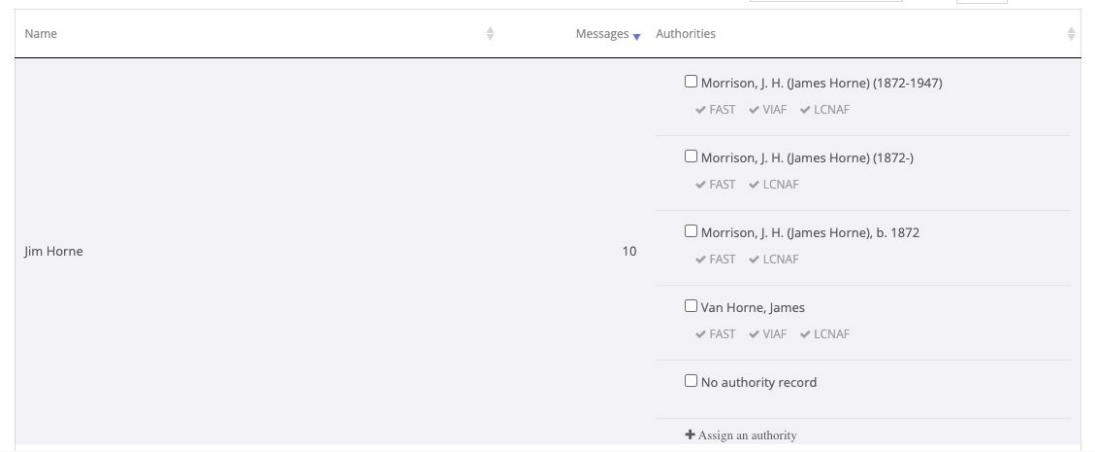


Switching modules

- Edit the module in the epadd.settings file
  - e.g. epadd.mode=processing
- Switch the module in the browser - http://localhost:9099/epadd/switch-module
  - Appraisal → Processing → Delivery/Discovery 👍
  - Delivery/Discovery → Appraisal/Processing 👎

Reviewing and confirming authorities

- Feature only available in the Processing Module
- Reconcile correspondents and named entities extracted by ePADD with authorized headings
- OCLC, FAST, VIAF, LCNAF



ePADD data structure

- When an account is imported into ePADD (whether from a file or via IMAP), ePADD generates a new set of data to represent it in the application
- Any changes made to a collection in ePADD do not alter the original .mbox file that was imported

Managing ePADD data

- ePADD collection data is stored at:
  - c:\users\<username>\ [Windows]
  - HD/Users/<username> [Mac]
- Each module has a corresponding folder:
  - epadd-appraisal (created automatically)
  - epadd-processing (created manually by user)
  - epadd-delivery (created manually by user)
  - epadd-discovery(created manually by user)
- Each module generates a unique package for that module:
  - Appraisal → “ePADD archive of _____”
  - Processing → “ePADD archive of _____”
  - Delivery → “ePADD archive of _____-Delivery”
  - Discovery → “ePADD archive of _____-Discovery”

Processing tips and strategies

- Start by browsing – get to know the collection!
- Consider the context and research interest for your collection and how this informs your processing strategy
- Track commonly restricted terms and correspondents
- Customizing lexicons
  - Create new lexicons based on the collection or with input from the donor/creator or curator
  - Remove terms that cause false positives

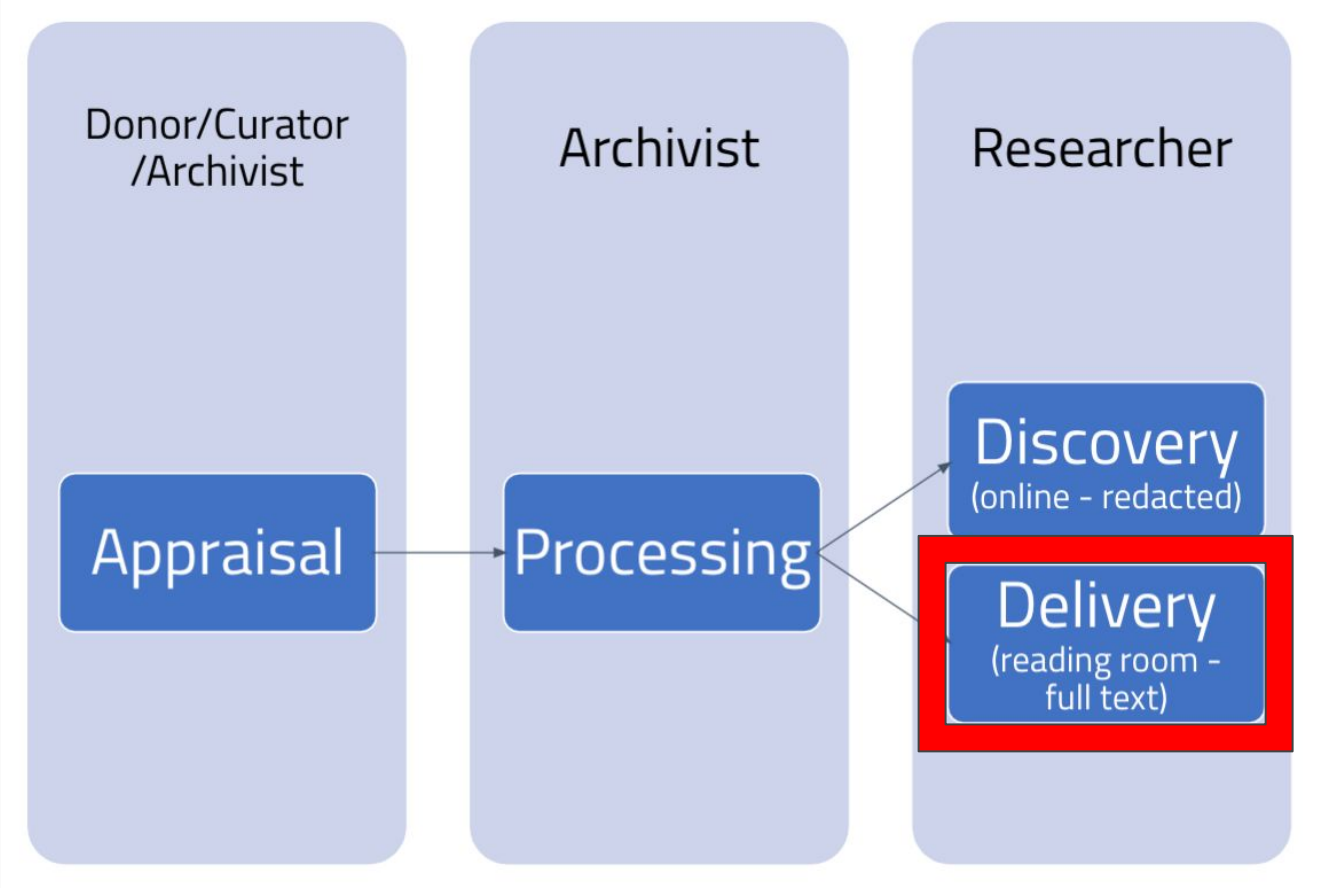
Launching ePADD from the command line

- Allows users to select more RAM than default to run ePADD
- Must use epadd-standalone.jar file
- How to:
  - Open Terminal or Command Line
  - Use “cd” command to navigate to directory where the epadd-standalone.jar file is stored
  - Enter the command: java -Xmx#g -jar epadd-standalone.jar

    *replacing the “#” with the amount of RAM that you would like to allocate

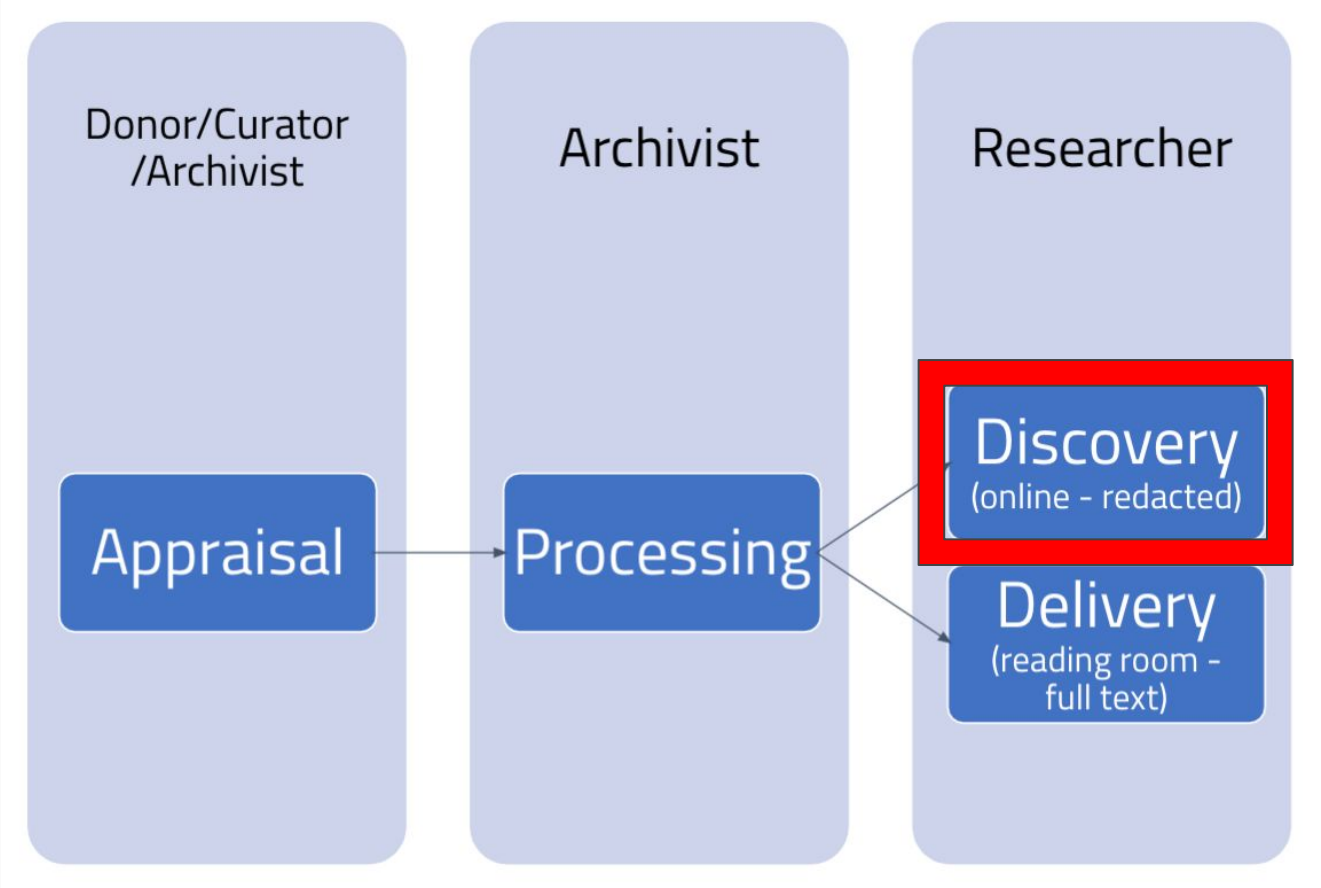
Quitting ePADD

- Two ways to quit ePADD
  - Quit from the ePADD icon in the top bar (Mac) or bottom right corner (Windows)
  - Quit from the Command Line by entering the ctrl+c command


Donor/Curator /Archivist
Archivist
Researcher
Appraisal
Processing
Discovery
(online - redacted)
Delivery
(reading room - full text)

Delivery module

- The ePADD module that is installed in an institution’s reading room
- Includes full text and attachments of all messages that have not been restricted
- Same searching options as in Appraisal/Processing Modules
- Users can label and annotate messages to aid research
- Users can download selected messages if allowed your institutional policies


Donor/Curator /Archivist
Archivist
Researcher
Appraisal
Processing
Discovery
(online - redacted)
Delivery
(reading room - full text)

Discovery module

- The (optional) ePADD module that can be hosted online to facilitate discovery of email collections
- Displays a redacted version of the email collection to protect privacy of individuals in collection
- Users cannot label or annotate messages

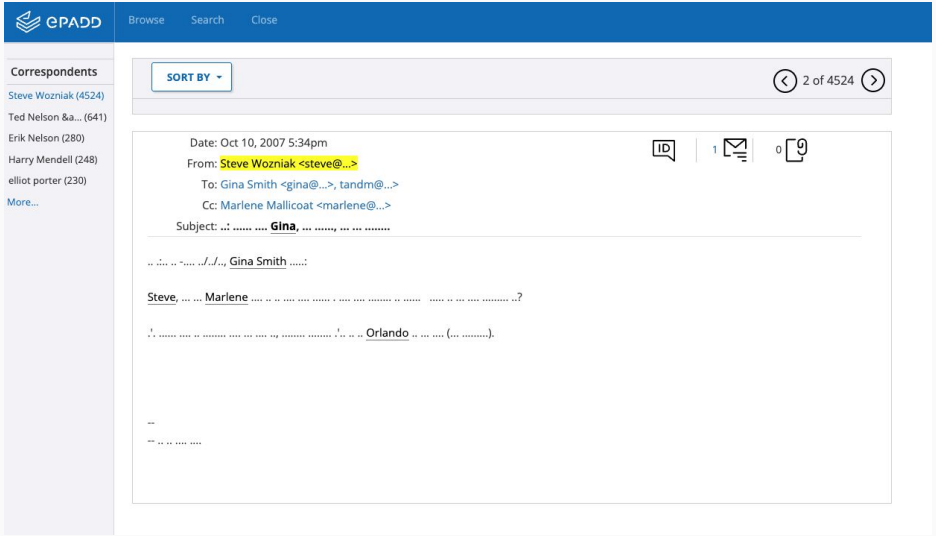
Discovery module

What is there?

- Collection metadata (similar to finding aid frontmatter)
- Redacted email headers
- Names of correspondents
- Extracted entities

What is *not* there?

- Full text of messages
- Full email addresses of correspondents
- Attachments



Shared discovery module

- Shared version of Discovery Module that any institution can contribute to
- Hosted at epadd.org
- How to contribute:
  - Contact us at epadd_project@stanford.edu
  - Check out the ePADD Shared Discovery Module Collection Contributors user guide: https://bit.ly/3uNpUcN

How can you participate?

- Share your work and experiences
- Join the Steering Committee
- Join the Code Group
- Software testing
- Cross-Project Pollination

Resources for more information

- ePADD User Guides:
  - Main user guide -- https://bit.ly/3qmnVLV
  - Shared Discovery Module contributors guide -- https://bit.ly/3uNpUcN
- Email Processing and Preservation Interest Group -- https://stanford.io/3A1udVM

Thank you!

debauche@stanford.edu

- Visit **epaddproject.org**
- Follow **@e_padd**
- Watch **youtu.be/vu1Oi8TiGiU**
- Receive **epadd_list@stanford.lists.edu**
- Download / Contribute **github.com/epadd**
- Reach **epadd_project@stanford.edu**