

***Email Preservation Together:
The ePADD Open-Source
Software Community***

DPC Clinic



May 13, 2026

What we'll cover today

- Introduction to ePADD
- Shared Discovery Module
- Searching for sensitive information
- Redacting messages
- Experiments with AI
- Project sustainability
- Wrap up & resources

ePADD

Email Processing, Appraisal, Discovery, Delivery (and Preservation!)

ePADD is free and open source software developed by Stanford University's Special Collections & University Archives that supports the appraisal, processing, preservation, discovery, and delivery of historical email archives. ePADD incorporates techniques from computer science and computational linguistics, including machine learning, natural language processing, and named entity recognition to help users access and search email collections of historical and cultural value.



[Download and Install](#)



[User Guide](#)



[Emailchemy for ePADD](#)



[Get Involved](#)

Introduction to ePADD

Project Funders



NATIONAL
ARCHIVES

NATIONAL HISTORICAL
PUBLICATIONS
& RECORDS COMMISSION



Stanford
LIBRARIES

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

Smithsonian Libraries and Archives

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

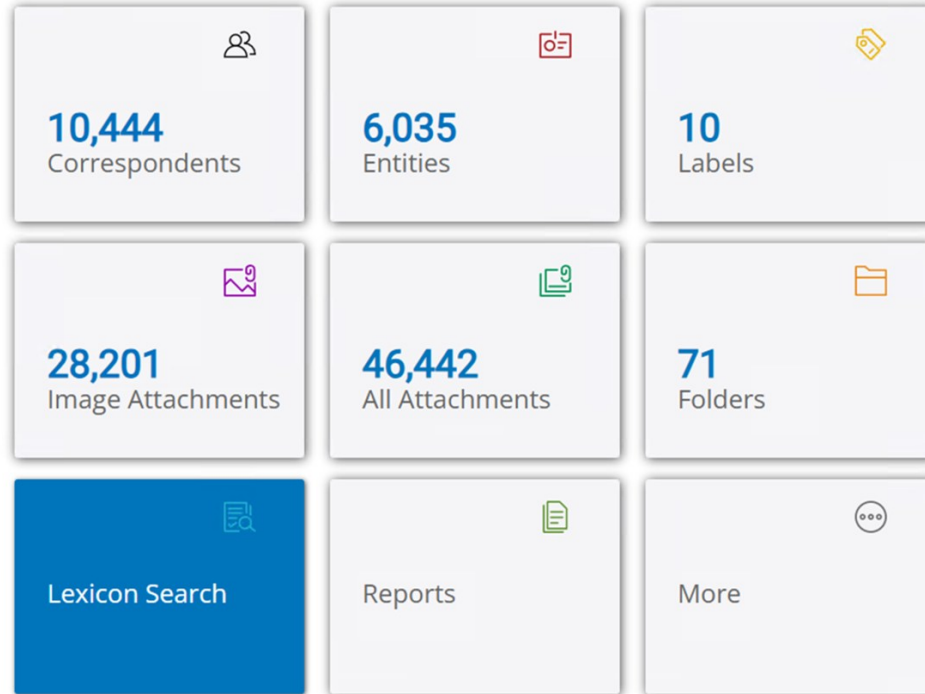
University of Virginia

University of Warwick

Wildlife Conservation Society

92NY, NYC

Searching for Sensitive Information



Searching for Sensitive Information

Lexicon	Number of Categories
persona.environmental.artist.projects.stanford	50
sentiments	42
sensitive	9
persona.academic.administrator.sensitive.duke	8
sensitive-ccca-german	8
persona.composer.nypl	6
persona.faculty.uci	4
persona.writer.theater.nypl	4
sensitive-cca-french	4
persona.author.princeton	3
persona.microbiologist.uiuc	3
persona.journalist.activist.politics.and.travel.ucb	2
regex	2

Searching for Sensitive Information



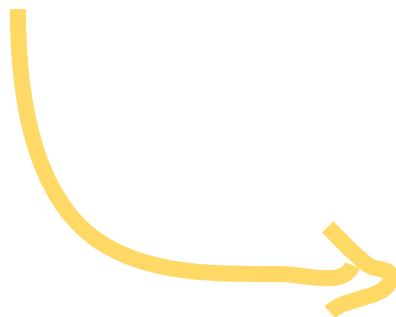
Search:

Lexicon category	Messages
Credit card number	118
U.S. social security number	10

Showing 1 to 2 of 2 entries

Credit card number	118
U.S. social security number	10

Showing 1 to 2 of 2 entries



Folder: [REDACTED]

Date: [REDACTED]

Labels: [REDACTED]

From: [REDACTED]

To: [REDACTED]

Subject: RE: contract & forms

The answers to 5-7 are correct (i.e. NO). SSN = [REDACTED]

[Megan](#)

LABELS ▾ SORT BY ▾ ATTACH

Restriction Labels

Do not Transfer

Transfer to Delivery Or

Do not Transfer

LABELS ▾

SORT BY ▾

ATTACHMENT VIEW



1 of 4



Do not Transfer

General Labels

Error in Attachments

Reviewed

Cleared for Release

No Date

Possibly Bad date

Other error while Parsi

Error in Corresponden

Permission Labels

Meg Permitted for transfer



Folder: [REDACTED]

Date: Jan 22, 2013 9:51am

Labels:

From: [REDACTED]

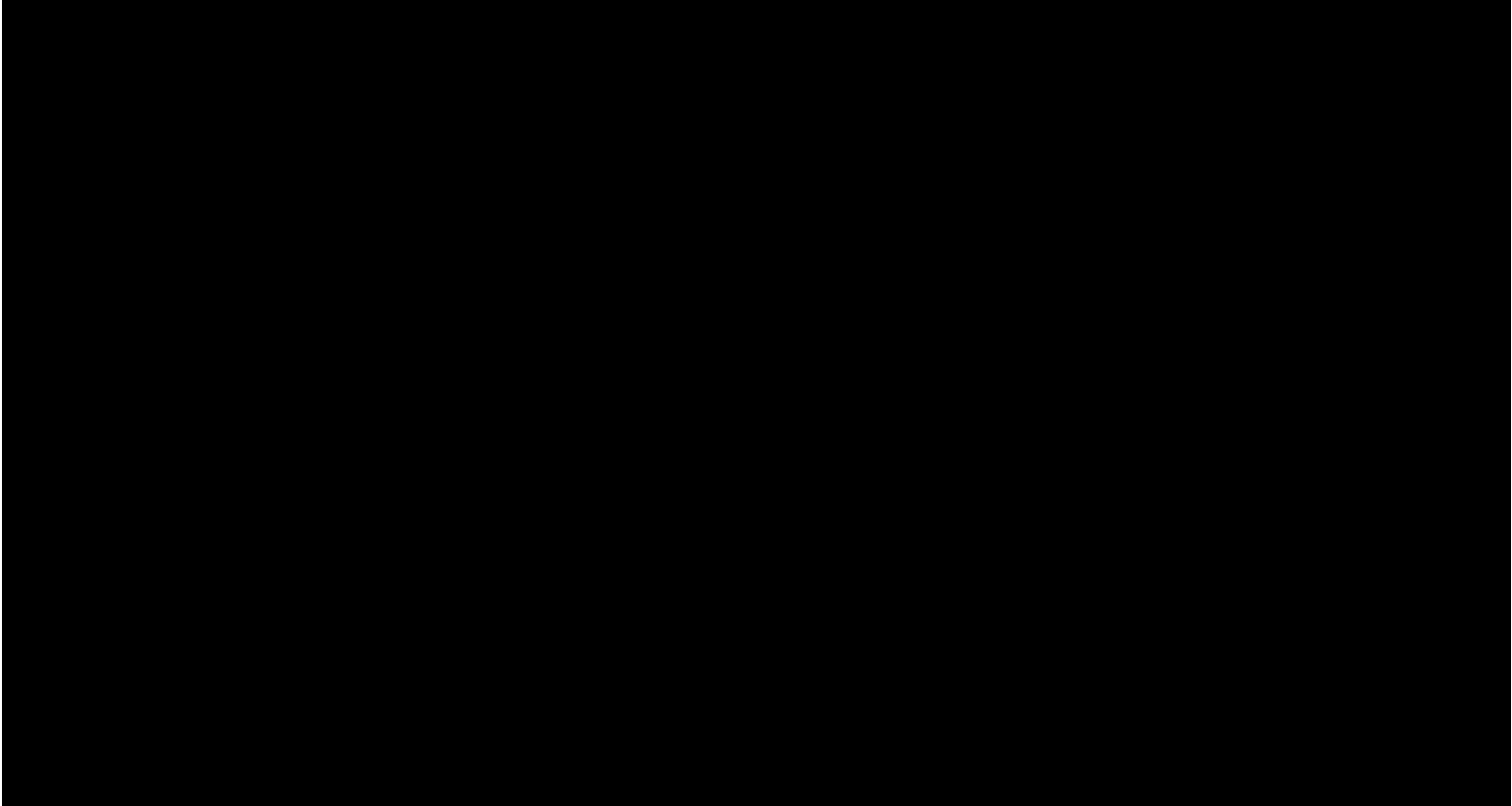
To: [REDACTED]

Subject: [REDACTED]

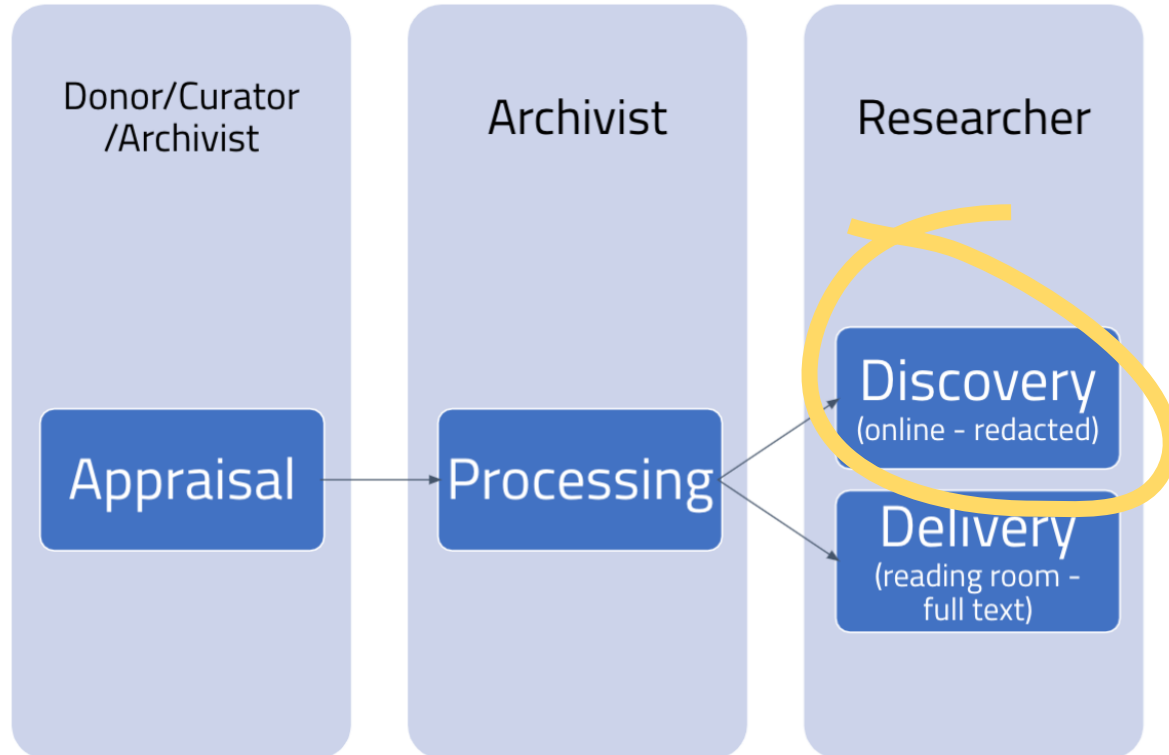
The answers to 5-7 are correct (i.e. NO). SSN = [REDACTED]

Megan

Message Redaction - Demo



Shared Discovery Module



Shared 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 (not even filenames)

The screenshot displays the EPADD Shared Discovery Module interface. The top navigation bar includes 'Browse', 'Search', and 'Close' options. A sidebar on the left lists 'Correspondents' with names like Josh Harrison, Elizabeth Tho., Katherine Pau., NYTimes.com, and Bill Fox, along with a 'More...' link. Below this is a 'Sender' section for 'Owner (10075)'. The main content area shows an email header with a 'SORT BY' dropdown, a page indicator '5 of 65084', and icons for ID, email, and a document. The email details are: Date: Sep 8, 2004 11:23am; From: Jeroen van Westen <jvanwesten@...> (redacted); To: Helen and Newton Mayer Harrison <harrstudio@...> (redacted); Subject: Potsdam - Alberto Duman. The email body is mostly redacted with a large grey block, but some text is visible: 'Helen and Newton', 'Diane Karp', 'New Mexico', 'Gerco de Ruijter and Anne Ausloos, And.', 'Santa Fe', 'Cardiff', 'Alberto Duman', 'Potsdam Conference', and 'Jeroen'.

Shared Discovery Module

Shared version of Discovery Module that any institution can contribute to

- Hosted at epadd.stanford.edu
- How to contribute:
 - Contact us at epadd_project@stanford.edu
 - out the ePADD Shared Discovery Module Collection Contributors user guide: <https://bit.ly/3uNpUcN>

AI for Email Appraisal and Analysis

Plug in your model of choice

Local Setup (Llama):

- Dedicated library PC with powerful GPU
 - Running Ollama for using local LLMs
-

- Pros:
 - No data sent to third parties
 - Full control over models
- Cons:
 - Requires local hardware & maintenance
 - May have lower performance compared to large cloud models

Cloud Setup (ChatGPT):


- ePADD sends data to cloud API
-

- Pros:
 - Access to powerful, latest models
 - No local hardware required
 - Easy to scale
- Cons:
 - Data leaves your institution
 - Potential privacy & sensitivity concerns
 - API costs might occur

About this Archive

 Name of Archive Owner

John Doe

 Primary Email Address

 Question to AI

What is the sentiment of each email from -1 to +1?

Possibly Bad date



Folder: D:\Jeb Bush1\JebEmails mboxfull\08 August 2002 Public.pst\August 2002 Public -2-\Top of Personal

Date: Aug 28, 1956 12:43am

Labels:

From: "Dr. Arnhilda Badia, PhD" <badiaa@fiu.edu>

To: jeb@jeb.org

Subject: **acto**

Jeb:

Tengo una idea. Podemos hacer una conferencia de prensa en la que tú estés con Carlos para presentar el cheque de \$100,000 que pasó en la Legislatura para comenzar el Jose Marti Scholarship Endowment. ¿Tú crees que puedas venir antes del día 10 de septiembre? Avísame para organizar el acto.

Arnhilda

Annotation



Score: +0.6 Reason: The message is constructive and upbeat, proposing a collaborative public event around funding and coordination, with a polite and forward-looking tone.

Mildly offensive attachment



Folder: b.eml

Date: Jan 29, 2026 5:06pm

Labels:

From: [Weird Kid Software <emailchemydemo@weirdkid.com>](mailto:emailchemydemo@weirdkid.com)

To: [Jochen Farwer <jochen.farwer@manchester.ac.uk>](mailto:jochen.farwer@manchester.ac.uk)

Subject: **Please Purchase Emailchemy**

Annotation



The image contains a classical statue with partial nudity; however, it is recognized as a culturally significant artwork in a public setting. No sexual or offensive intent is detected. The content is safe.



[Jochen Farwer](#) | [Software Developer](#) | Red M.6 ([Main Library](#)) | [The University of Manchester Library](#) | [Oxford Road](#) | Manchester | [M13 9PP](#) |



Limitations & Next Steps for AI Integration

Limitations

- Large email archives take a long time to analyse

Next Steps

- Allow users to analyse a subset of emails within an archive
- Make model integration (e.g. ChatGPT) user-friendly

ePADD and AI tools for Appraisal



Challenges

- Sensitivity review and regulatory environment
- Volume of records to be appraised



Appraisal Experience

- Manual individual appraisal - 1 week to review and label approximately 1300 email messages
- Can AI for an initial classification speed things up?

Sentiment and Medical Data Analysis

Annotation  



response to the decision not to intervene surgically. -----
Medical content score: 0.75. The email mentions a medical decision not to intervene surgically and a scheduled scan in six months, indicating a moderate level of medical context.

Annotation  



Sentiment score: 0.9. The tone is overwhelmingly positive and inviting due to the enthusiastic language used in extending an invitation. -----
-----Medical content score: 0.0. No medical context found in the email.

Label	Type	Messages
Possibly neutral sentiment (score between -0.6 and +0.6)	General	487
Possibly very positive sentiment (score > 0.8)	General	447
Possibly containing medical information	General	254
Possibly positive sentiment (score > 0.6)	General	136
Possibly negative sentiment (score < -0.6)	General	33
Possibly Bad date	General	0
Possibly very negative sentiment (score < 0.8)	General	0

Application of Data Protection Act

Annotation  

Sensitivity score base on data protection act: 0.0. The email contains personal and medical information but it is a private communication between two individuals with no evidence of data protection breach or misuse.

Annotation  

Sensitivity score base on data protection act: -0.1. The email does not contain any personal or sensitive information and appears to be a discussion about poetry, but it does mention specific individuals which could potentially be considered as processing of personal data.

Neutral sensitivity score X	General	766
Error in Correspondents	General	19
Very neg. sensitivity score X	General	10
Very pos. sensitivity score X	General	8
neg. sensitivity score X	General	7

Lessons Learned & Next Steps

Lessons Learned

- Value in exploring as a way forward for appraisal of large volumes of email
- Lack of consistency and reliability is still a risk

Next Steps

- Create a very detailed and specific appraisal rubric
- Thoughts and suggestions on AI for email appraisal and processing are welcome!

Project Sustainability

Previous model



Project needs



Current plan

- Development supported by grant funding, with ad hoc investments in maintenance and enhancements
- 2023: Code and Steering Groups launch; Dev work sustained through Manchester team (with some short-term investment from Harvard)
- Community contributions peppered throughout
- Resilience to personnel changes - support not anchored in a single person or supporting institution
- Assurance of basic maintenance, security, and stability
- Maintain open-source (and hopefully broaden, strengthen contributor community)
- Low to no cost, or flexible/optional costs
- Partnering with the Open Preservation Foundation (~Fall 2026 transition)
- Will address needs:
 - Maintain a stable code base
 - Strengthen our contributor model
 - Anchor our governance and broaden our base
- If you aren't already a member, consider joining OPF!

Get Involved with ePADD

- Eager to support the sustainability of ePADD?
 - Join the Steering Group: <https://www.epaddproject.org/community/steering-group>
- Interested in making technical contributions to ePADD?
 - Join the Code Group: <https://www.epaddproject.org/community/code-group>
- Want to learn more about email preservation?
 - DPC Technology Watch: Preserving Email - <http://doi.org/10.7207/twgn21-08>
 - DPC Novice-to-Know How: <https://dpc.getlearnworlds.com/course/n2kh-email>
- Curious about seeing ePADD in action?
 - How-to videos: <https://www.epaddproject.org/resources/videos>
- Ready to get started using ePADD?
 - Download the software on Github: <https://github.com/ePADD/epadd>
- Already using ePADD and want to talk with others?
 - Join the ePADD Google Group: <https://groups.google.com/g/epaddusers>