Finding light in dark archives: Using AI to connect context and content in email

LUSTRE (Unlocking our Digital Past with Artificial Intelligence)

Workshop One

AI and born-digital archives: Challenges and opportunities

Thursday, 26th January 2023

Dr Adam Nix & Prof Stephanie Decker, University of Birmingham

AI & SOCIETY
https://doi.org/10.1007/s00146-021-01369-9

ORIGINAL ARTICLE

Finding light in dark archives: using AI to connect context and content in email

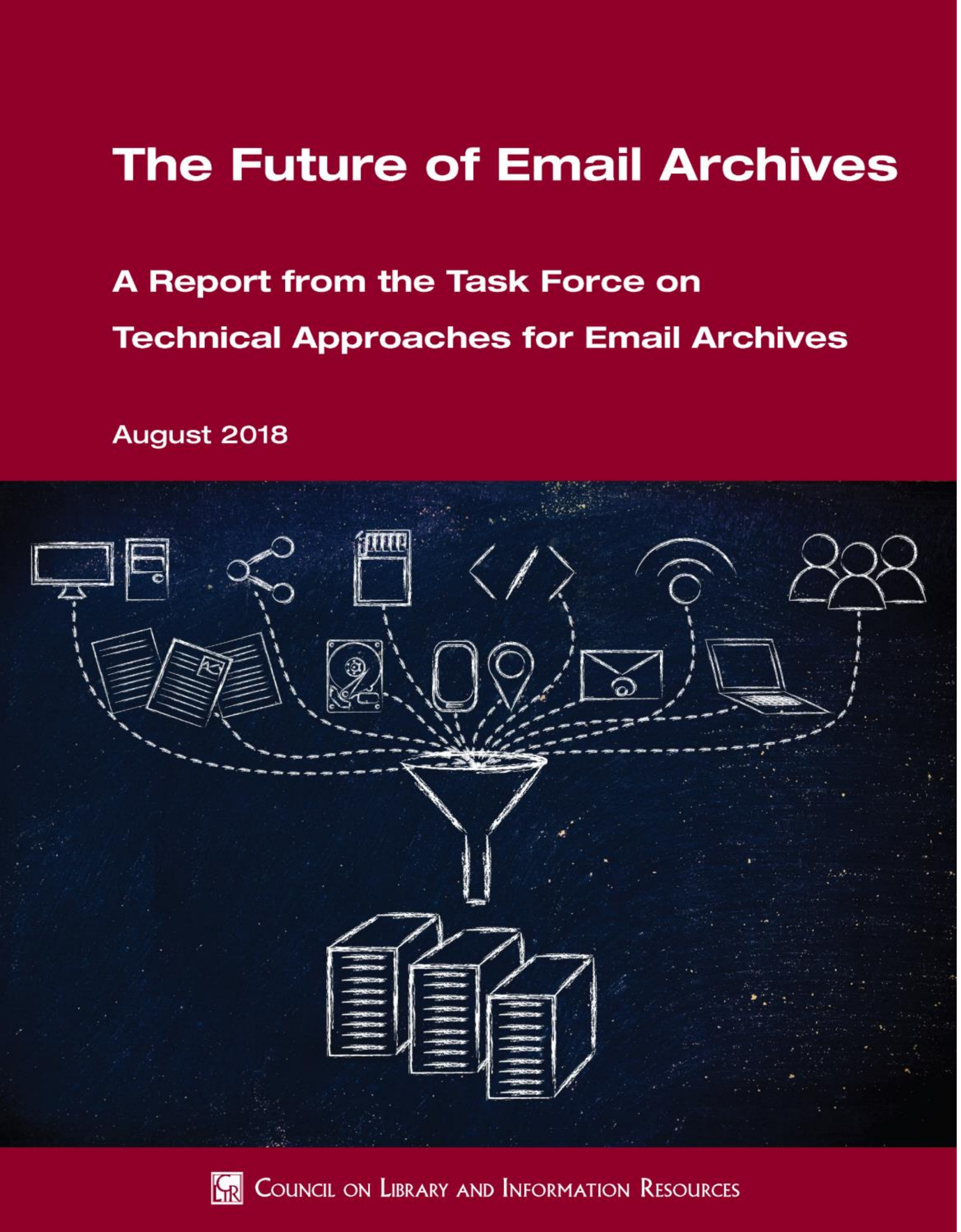
Stephanie Decker[1] · David A. Kirsch[2] · Santhilata Kuppili Venkata[3] · Adam Nix[4]

Received: 29 April 2021 / Accepted: 23 November 2021

Abstract

Email archives are important historical resources, but access to such data poses a unique archival challenge and many born-digital collections remain dark, while questions of how they should be effectively made available remain. This paper contributes to the growing interest in preserving access to email by addressing the needs of users, in readiness for when such collections become more widely available. We argue that for the content of email to be meaningfully accessed, the context of email must form part of this access. In exploring this idea, we focus on discovery within large, multi-custodian archives of organisational email, where emails' network features are particularly apparent. We introduce our prototype search tool, which uses AI-based methods to support user-driven exploration of email. Specifically, we integrate two distinct AI models that generate systematically different types of results, one based upon simple, phrase-matching and the other upon more complex, BERT embeddings. Together, these provide a new pathway to contextual discovery that accounts for the diversity of future archival users, their interests and level of experience.

Keywords Email archives · Born-digital collections · Computational archival studies · Contextual email discovery



The Future of Email Archives

A Report from the Task Force on Technical Approaches for Email Archives

August 2018

Council on Library and Information Resources

"Electronic records in archival repositories, especially email messages, are fundamentally different. Traditional paper-based series of correspondence are often uniform in their contents and structure, whereas email collections include both formal and informal communications, mass mailings from listservs, and even unsolicited advertising that, when combined with the volume of messages, makes traditional records management difficult if not impossible"

Born-digital access and discovery

- Many born-digital collections remain inaccessible or ‘dark’ while access issues are negotiated.
- Those already available can be difficult to search, particularly qualitatively.
- Ethical and privacy implications of large digital collections are serious but also unclear.
- Answering the problem of access requires collaboration ‘between both sides of the reading room’(Jaillant, 2019)

Building a born-digital user perspective

- Born-digital sources are vital for researchers of the post-analogue past
- For many, scale and complexity are going to challenge predominantly analogue assumptions
- They also offer useful affordances and potential for new insights:

New perspectives

- Marginalized actors
- Day-to-day life

Contextual specificity

- Dating and timing
- Audience and reach

Advanced searchability

- Scaled searches
- Replicability

(Nix and Decker, 2021)

Different users, different uses

- Some uses want complete datasets in their original form
- Many users expect some curation, and are still relevantly inexperienced (Wellcome Trust, 2017)
- Talboom and Underdown (2019) identify three user types:

Reader

- Wants to access a digital source like a traditional paper source

Digitally Curious

- Wants to search large databases to identify items of importance for more in-depth study

Data User

- Wants to perform computational analysis over entire collections

How will users *actually* engage with born-digital archives once access issues have been navigated?

Email specific issues

- The networked nature of email is key to meaningful access:
  - Email is a hybrid artifact: email IS and email ARE
  - Not just information as content, but also as context
- Scholars often focus on contextual aspects over content:
  - Frequency and networks (Aven, 2015)
  - Timing and sequencing (Byun and Kirsch, 2020)
  - Language (Wright, 2013)
- However, for content to be meaningful, both individual and network aspects of email need to be maintained.
  - i.e., context and content

Aims & Scope of our Research Projects

Outline an historical-use perspective on born-digital sources like email

Reflect on born-digital discovery, use artificial intelligence and email as a source

Describe our approach to discovery, connecting context and content within organizational email

Contextualizing Email Archives Project

Explored new ways to make email archives available to search and study while maintaining the relational and network properties of the format

- Used the emails of a failed US Dot-Com company
- Preserved and made available for research via the LDC
- Collaboration with TNA digital archives specialists

Created a digital history telling the story of the company, based primarily on its email archive.

Making email accessible as historical source

https://dotcomarchive.
bristol.ac.uk/

Contextualising Email Archives

Hi, I'm Mr Dummy.

Please click on the computer screen to begin the journey.

Round One

University of Chicago, Library $40,230.00	• Attachment Converter • Streamline the conversion of email attachments
University of Albany, SUNY $63,890.00	• Mailbag • Introduces a near-to-capture packaging
Columbia University $98,630.04	• Document Cloud
Council of State Archivists, Inc. $100,000.00	• Provide capacity building activities to state and territorial archives
Harvard University $100,000.00	• ePADD+ • Enhance processing and preservation capabilities

Round Two

University of Maryland $56,949.96	• EMCODIST
University of North Carolina $87,716.81	• RATOM-FIRE
92nd Street Y $100,000.00	• ePADD

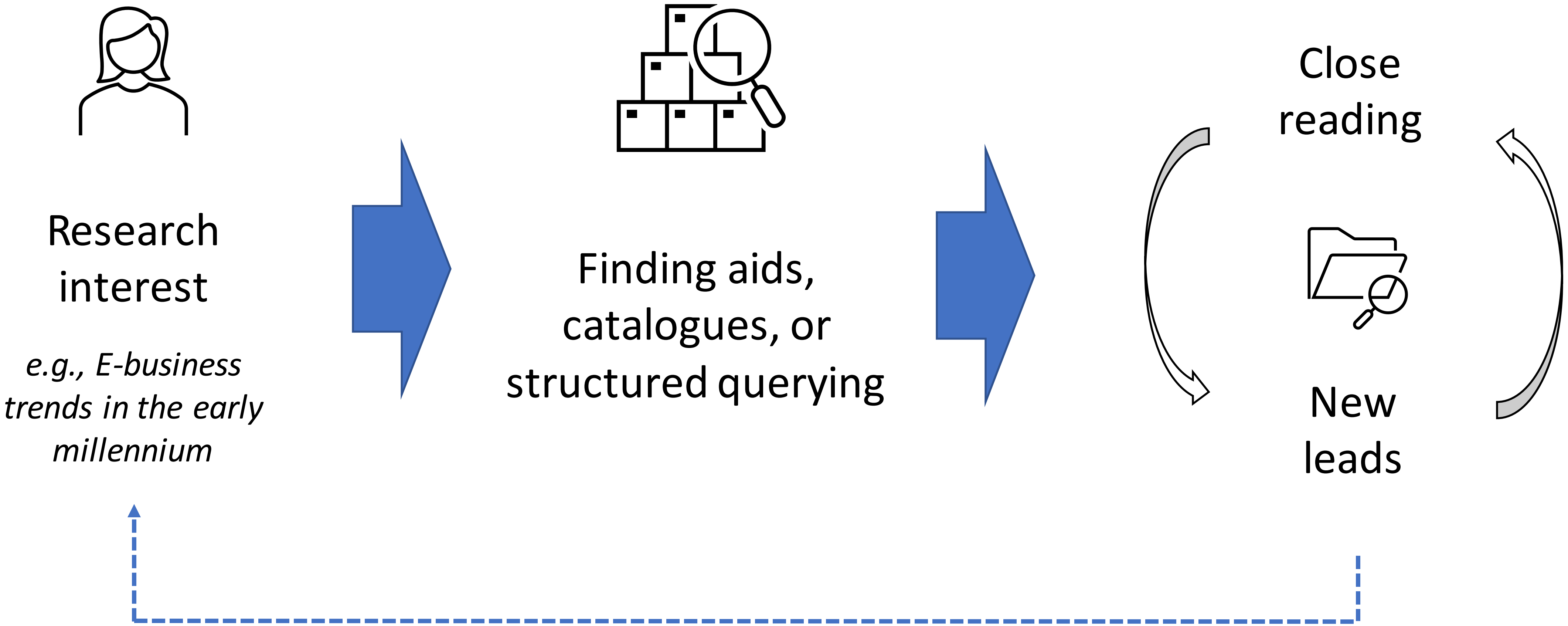
Working Hypothesis

With a relatively complete e-mail archive, a scholar can ask and answer most important historical questions about an organization.

Methodological Problem

While emails offer valuable insight, a lack of context often presents challenges to those wishing to qualitatively understand their content, inter-relationship, and wider historical and theoretical significance.

Traditional Archival Discovery



Research interest

e.g., E-business trends in the early millennium

Finding aids, catalogues, or structured querying

Close reading

New leads

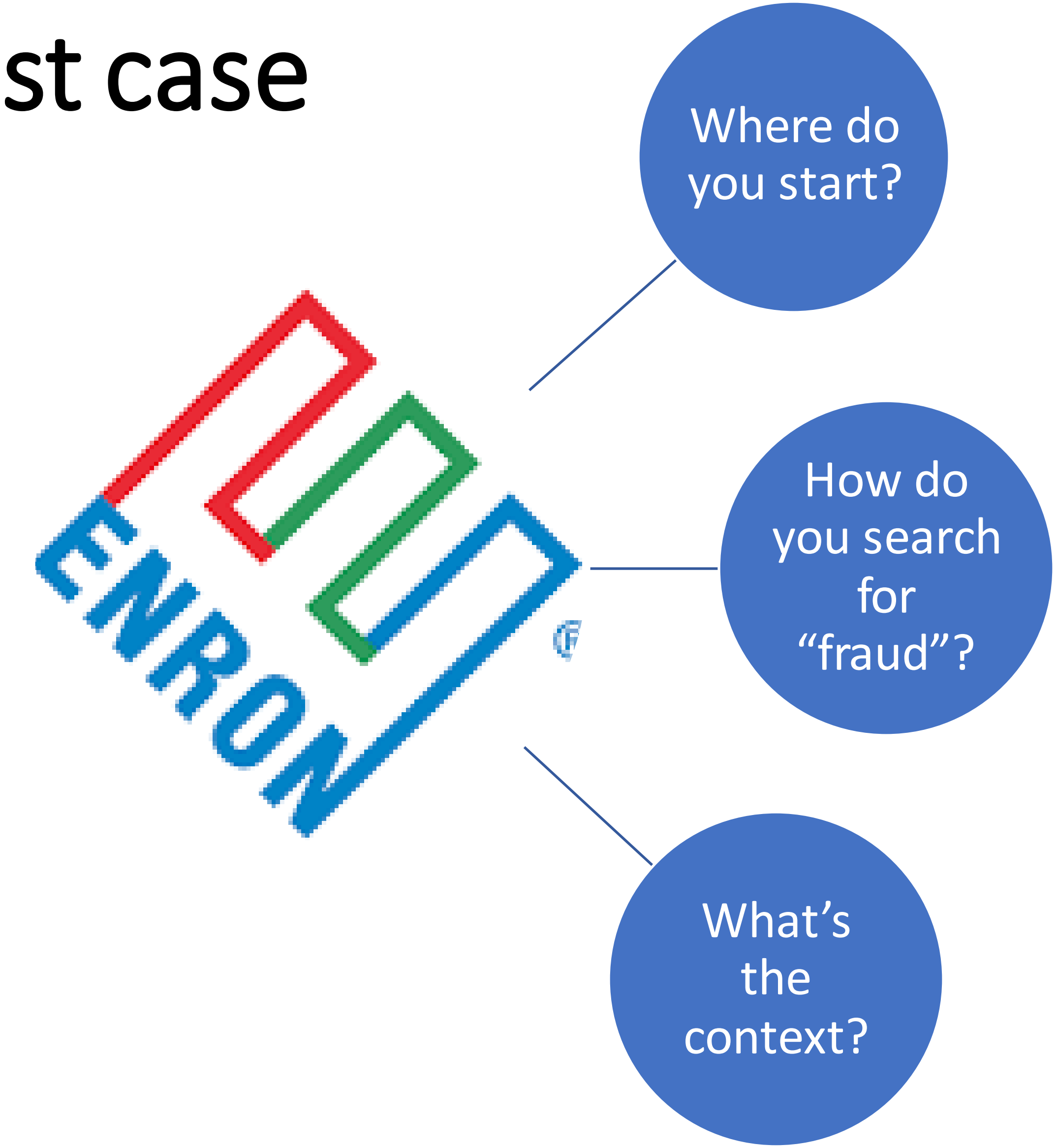
How do users navigate the empty search box?

Enron: the EMCODIST test case

The Enron Email Corpus was made public in the mid-2000s by the FERC.

Contents has been widely used by computer scientists to understand email behavior

Seen more limited usage in social scientific research
(e.g., Aven, 2015; Benke 2018; Nix et al, 2021)



Where do you start?

How do you search for “fraud”?

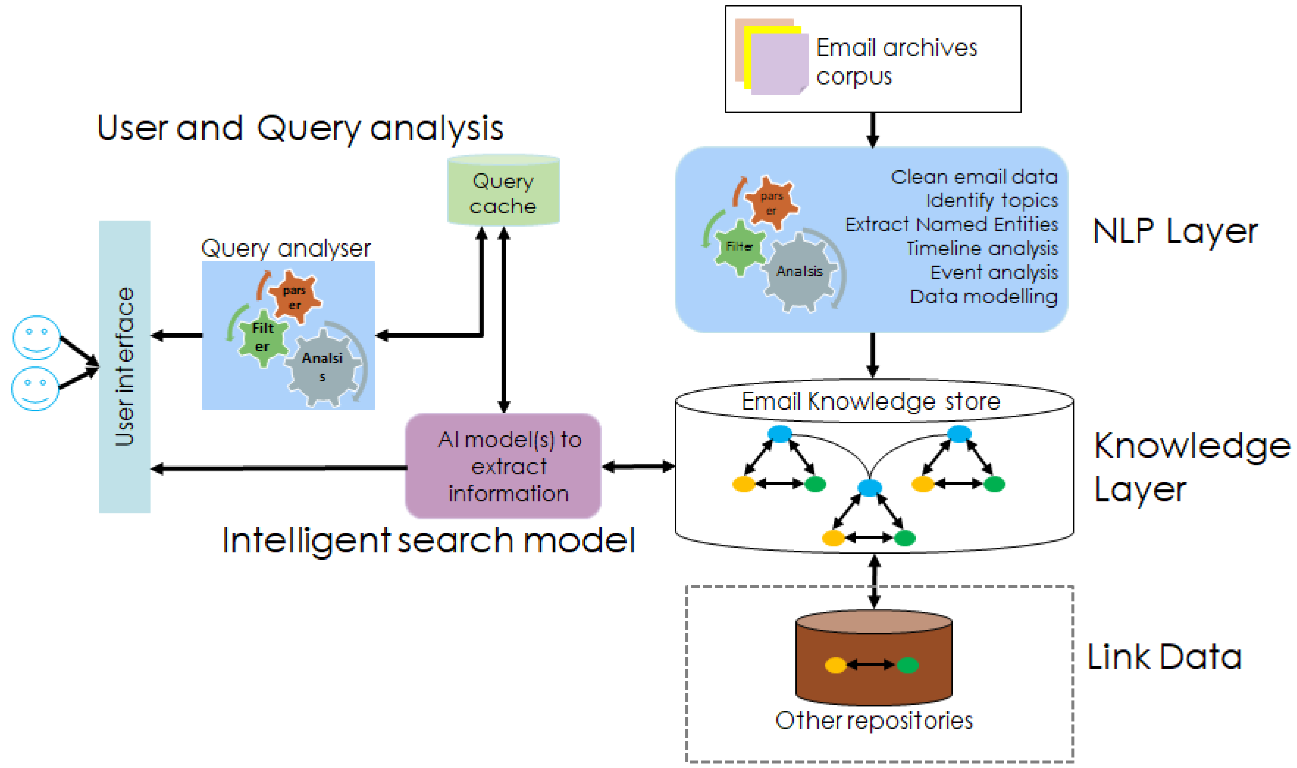
What’s the context?

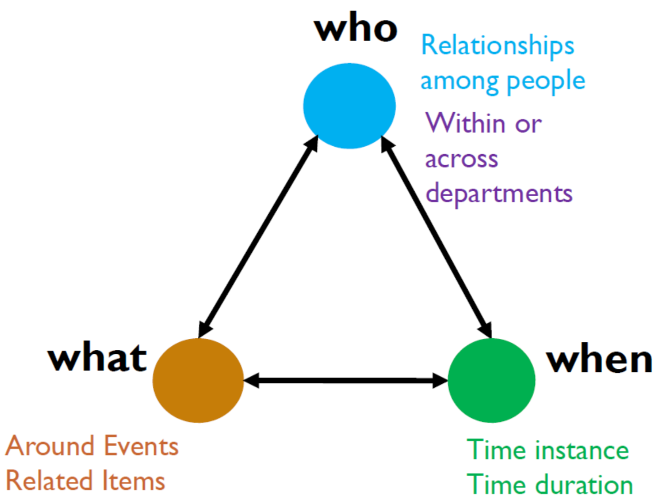
EMCODIST code on GitHub: https://github.com/Contextualising-Email-Archives/discovery-tool

Aims of EMCODIST

- Finding out how researchers actually search email corpora
- Build an understanding of user preferences for email as data
- Contribute to the development of best practices for email-based research
- Develop a tool that allows scholars to search, read and analyse large email datasets

Developing a tool for email archive search and discovery

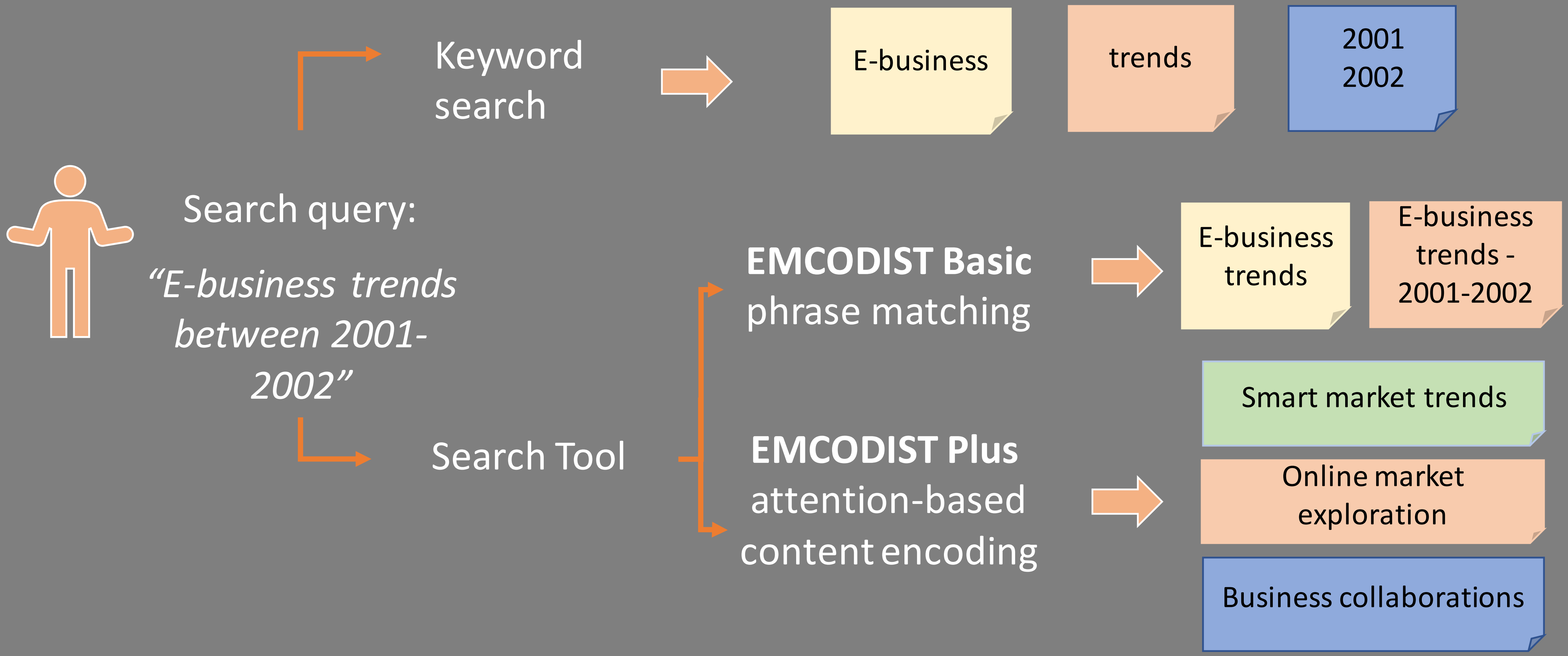




Requirements for effective machine-assisted discovery:

- Search phrases using natural language understanding
- Restrict scope based on contextual factors (time, user, message type)
- Allow connections across contextually relevant locations (different email accounts)

EMCODIST Plus



Search query:

"E-business trends between 2001-2002"

Keyword search

E-business

trends

2001
2002

Search Tool

EMCODIST Basic phrase matching

E-business trends

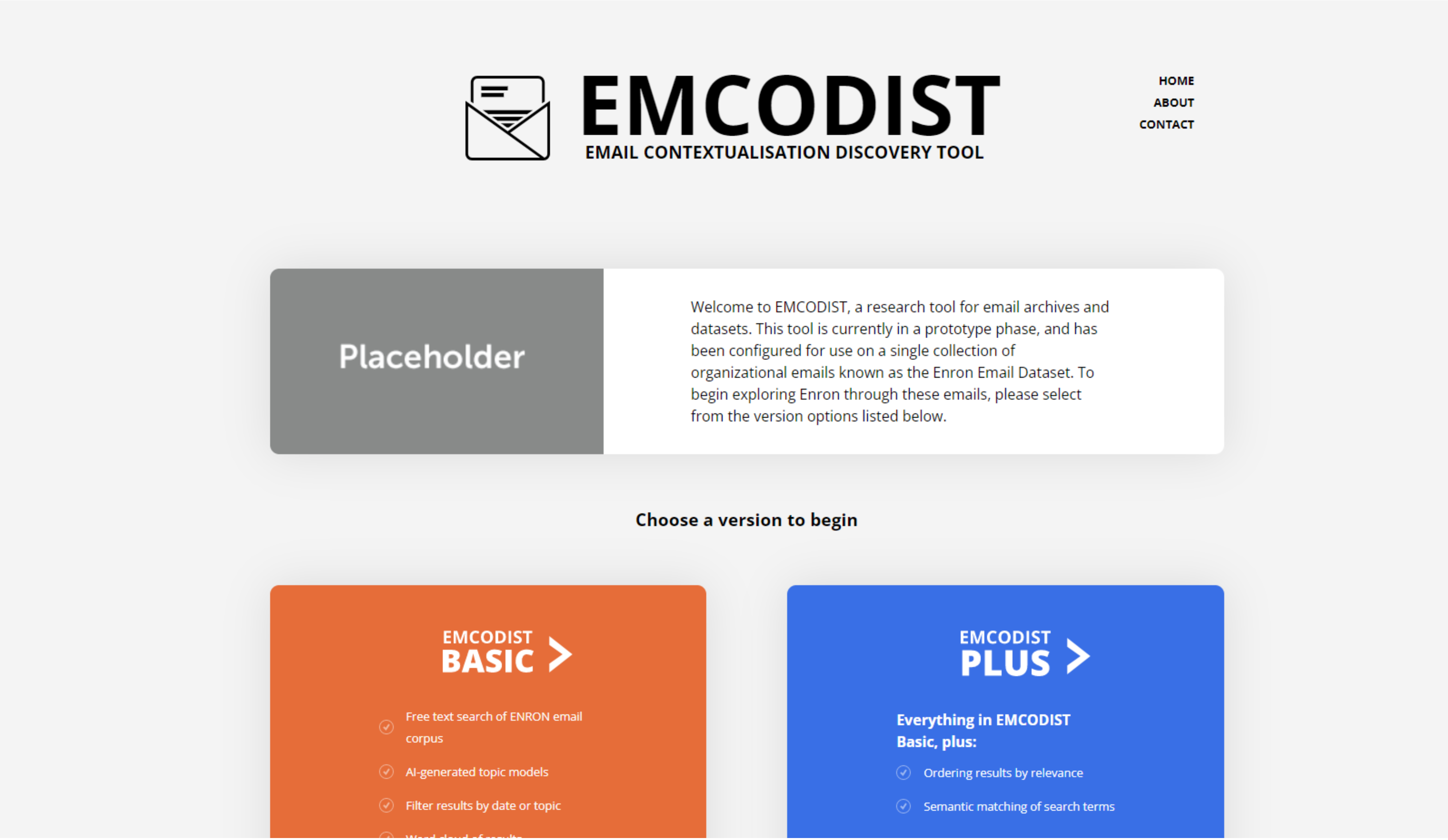
E-business trends - 2001-2002

EMCODIST Plus attention-based content encoding

Smart market trends

Online market exploration

Business collaborations



EMCODIST

EMAIL CONTEXTUALISATION DISCOVERY TOOL

Welcome to EMCODIST, a research tool for email archives and datasets. This tool is currently in a prototype phase, and has been configured for use on a single collection of organizational emails known as the Enron Email Dataset. To begin exploring Enron through these emails, please select from the version options listed below.

Choose a version to begin

EMCODIST BASIC >

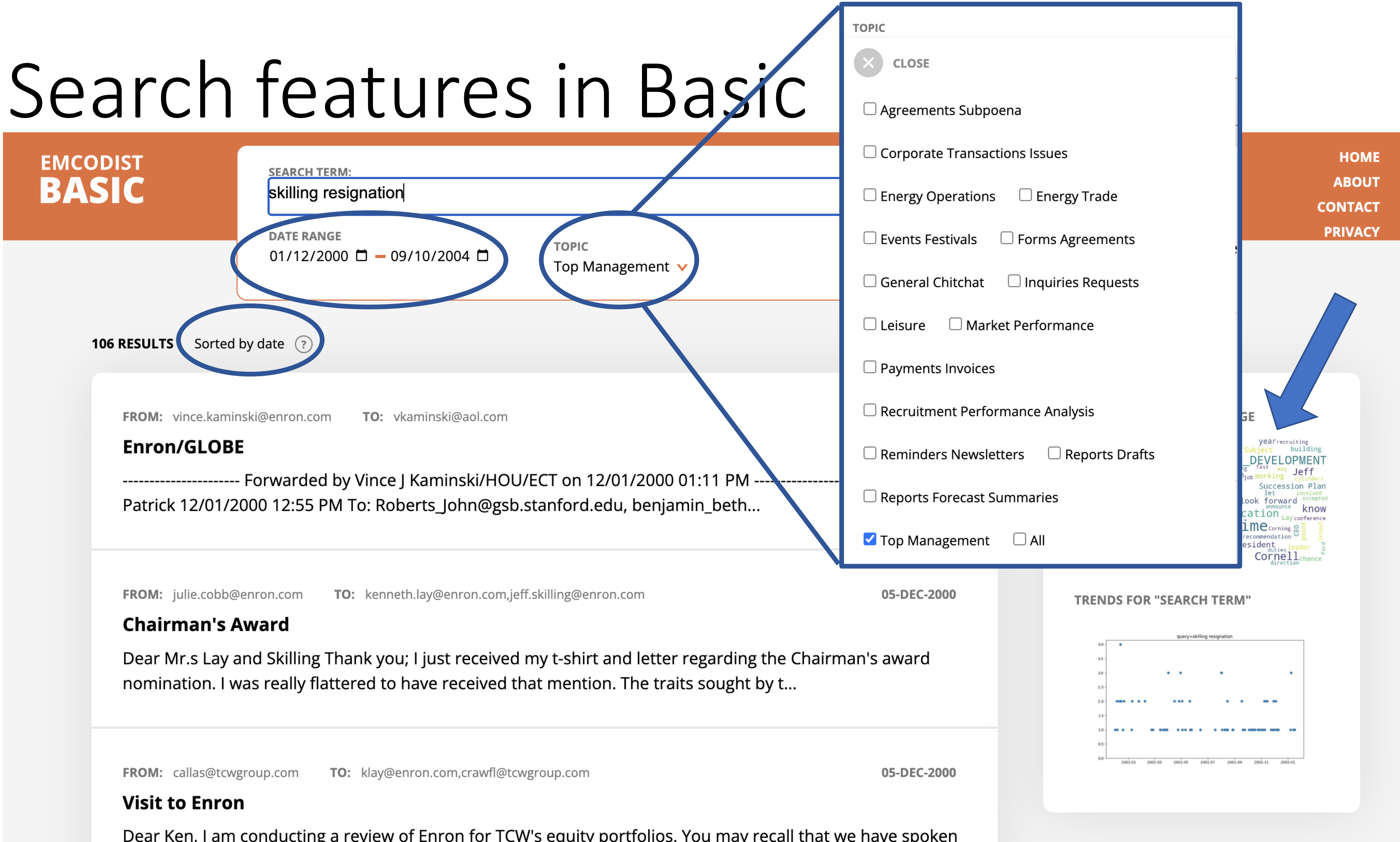
- ✓ Free text search of ENRON email corpus
- ✓ AI-generated topic models
- ✓ Filter results by date or topic
- ✓ Word cloud of results

EMCODIST PLUS >

Everything in EMCODIST Basic, plus:

- ✓ Ordering results by relevance
- ✓ Semantic matching of search terms

Search features in Basic



SEARCH TERM: skilling resignation

DATE RANGE 01/12/2000 – 09/10/2004

TOPIC Top Management

TOPIC

CLOSE

☐ Agreements Subpoena

☐ Corporate Transactions Issues

☐ Energy Operations ☐ Energy Trade

☐ Events Festivals ☐ Forms Agreements

☐ General Chitchat ☐ Inquiries Requests

☐ Leisure ☐ Market Performance

☐ Payments Invoices

☐ Recruitment Performance Analysis

☐ Reminders Newsletters ☐ Reports Drafts

☐ Reports Forecast Summaries

☑ Top Management ☐ All

106 RESULTS Sorted by date ?

FROM: vince.kaminski@enron.com TO: vkaminski@aol.com

Enron/GLOBE

---------------------- Forwarded by Vince J Kaminski/HOU/ECT on 12/01/2000 01:11 PM ---------------
Patrick 12/01/2000 12:55 PM To: Roberts_John@gsb.stanford.edu, benjamin_beth...

FROM: julie.cobb@enron.com TO: kenneth.lay@enron.com,jeff.skilling@enron.com 05-DEC-2000

Chairman's Award

Dear Mr.s Lay and Skilling Thank you; I just received my t-shirt and letter regarding the Chairman's award nomination. I was really flattered to have received that mention. The traits sought by t...

FROM: callas@tcwgroup.com TO: klay@enron.com,crawfl@tcwgroup.com 05-DEC-2000

Visit to Enron

Dear Ken, I am conducting a review of Enron for TCW's equity portfolios. You may recall that we have spoken

Key features:

	EMCODIST Basic	EMCODIST Plus
Model Technique	Phrase matching by NLP techniques	BERT embeddings for document Classification and development of knowledge graph
Fairness & Bias	No bias observed as the model works on phrase matching	The bias induced by affiliation to entities in the context can be eliminated by training on large data volumes.
Error potential	Since this model looks for exact words, some of the resulting emails may not be relevant at all.	Even though best efforts are made to understand the context, some emails containing words with multiple meanings may be found in the result set.
Query complexity	Simple queries with specific keywords	Simple to medium complex queries (provide better results with more data)
Ideal user type	Knowledgeable users who have some idea about the contents of the corpus	Suggested for novice users without information about the corpus

Email archives should...

- Accommodate increasingly diverse research questions
- Enable users to work iteratively through a collection
- Fit with tacit & messy approaches to research
- Provide access for different levels of experience
- Offer relatively complete access to a whole organisational corpus

Thank you!

Contact: grp-email-archives@groups.bristol.ac.uk