Case Study: Developing AI search and retrieval tools to improve archival access using the AvocadoIT email collection

Kaila Fewster[1]

Educational applications: This case study is useful for exploring the challenges of email archives and preserving their context, along with exploring the use of Natural Language Processing (NLP) and deep learning techniques to support online search and retrieval. When it comes to search and retrieval, it is also illustrative of the importance of considering user-centred design when developing archival tools, but particularly AI-based archival tools. Additionally, this case study can be used to highlight some of the challenges associated with archiving born-digital records and how to meaningfully identify and encode their provenances within the metadata across collections.

Educational topics: AI for born-digital records, Natural Language Processing (NLP) in archives, HCI and HII with AI tools in archives, AI for access in archives, AI tool development in archives[2].

About: This case study is part of a series of learning materials developed by InterPARES Trust AI[3] researchers and educators to train archival professionals and students to effectively leverage artificial intelligence in their archival work. The final draft was completed on November 2nd, 2023.

This case study considers the challenges of making email and born-digital archives more accessible online. For the content of an email to be meaningfully accessed, its context must be preserved and also form part of this access. As such, this project aimed to create a prototype search tool that used AI methods to support user-driver exploration of an email archive.

Email archives are increasingly important historical resources; however, individual messages can be difficult to understand without broader contextualization about the organization, the individuals involved, and even the issue being discussed. In this sense, existing email archiving tools often fail to sufficiently retain this necessary contextual information, which makes

---

[1] InterPARES Trust AI Graduate Academic Assistant, University of British Columbia.

[2] Educational applications map to a Body of Knowledge proposed by InterPARES researchers for AI/ML for the archival professionals. https://docs.google.com/document/d/1UsjkkkGeSJrgCDJGASCAy5q0uo_ZkQpzi_Ch8XUcqYw/edit?usp=sharing

[3] This case study is an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia, and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). https://interparestrustai.org/

[4] Case Study: Using AI search and retrieval tools to improve archival access using the AvocadoIT email collection © 2024 by Fewster, Kaila is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

discovery difficult for users and archivists alike. Thus, the project team set out to develop a prototype search tool that connects the context of emails using knowledge extracted through Natural Language Processing (NLP). The dataset used to create this tool was an organizational email collection entitled AvocadoIT, licensed for scientific use by the Linguistic Data Consortium of the University of Pennsylvania. The data was collected from a now-defunct start-up from Silicon Valley in 2003, cleaned and assessed for risk through a government-funded grant, and made available for use in 2015.

In order to develop the search tool, two separate AI models were implemented. The project team envisioned email archive context as the interdependence of three main "entities": who, what and when, and used this definition to guide the development of the two AI models (See Figure 1). The first AI model analyses user queries through phrase searching to identify relevant names and events to determine the meaning and context of the question and pull up results with similar contexts. This model is ideal for context-knowledgeable users, who often already know what they are looking for and search with simple, descriptive queries. Alternatively, the second AI model uses a 3-step process of knowledge abstraction to define these contexts within the email corpus. This model is best for novice users, as it returns more relevant results with less precise search queries. The first step of this process is to extract the existing metadata properties from an email record. Then, using NLP, topics in the subject, body content and attachments are identified. Finally, the identified topics and metadata are linked to the recipient and sender to create a knowledge graph. The initial AI phrase-searching model then leverages this graph to reduce computational time when finding emails for a query (see Figure 2).

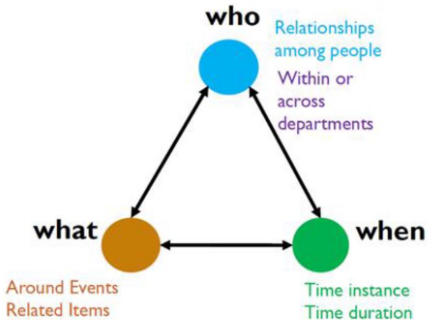

Figure 1: Context entities in email archives (Decker et al, 2022).

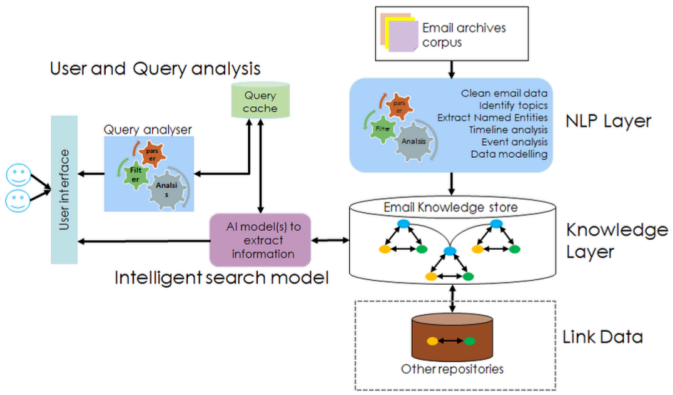

Figure 2: Overview of contextualisation discovery tool (Decker et al., 2022).

The second model uses an open-source neural network technique for NLP called Bi-directional Encoder Representation for Transformers (BERT) technology, which enables the model to understand a word's meaning in context with its neighbouring words. This allows for more dynamic and exploratory user searching, whereas the first phrase-searching model requires more focused searches. As such, to realize the full potential of email archives, it is necessary to pursue access in a way which reflects them as a born-digital medium and accounts for the various ways users will engage with them. While the landscape of digital archival discovery is expanding, ongoing support for contextualizing born-digital collections is necessary.

Possible Discussion Questions:

1. In what other ways can Natural Language Processing (NLP) and deep learning enhance search and retrieval functions in archives, and what limitations might this technology face in accurately interpreting and contextualizing archival content?
2. How do the challenges of archiving born-digital materials, such as email, differ from those associated with archiving physical materials, and how might AI applications be integrated into archival workflows to address these differences?
3. What other AI applications or models could improve the accessibility and usability of born-digital archives, and how might these tools evolve to support diverse user needs in the future?
4. What criteria should archivists and records managers use to evaluate the effectiveness of AI tools when considering to implement them into their workflows? How can this success then be meaningfully measured after integration?

References

Decker, S., Kirsch, D. A., Kuppili Venkata, S., & Nix, A. (2022). Finding light in dark archives: Using AI to connect context and content in email. *AI & Society*, 37(3), 859–872. https://doi.org/10.1007/s00146-021-01369-9

Decker, S., & Kuppili Venkata, S. (2023). *Contextualizing Email Archives* [Python]. Contextualising-Email-Archives. https://github.com/Contextualising-Email-Archives/discovery-tool (Original work published 2021)