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

1 Introduction

Email archives are important historical resources, but in contrast to pre-digital correspondence, an individual email message can be difficult to understand without proper contextual knowledge about the organization, the individuals involved, and the issues being discussed or debated. High flying investment banker Frank Quattrone was famously charged by the US Securities and Exchange Commission (SEC) for violating the law when, after being notified of an inquiry from the SEC, he forwarded his subordinates a message with the subject, "Time to clean up those files" and added the sentence, "I strongly advise you to follow these procedures." With the benefit of context at trial, a jury of Quattrone's peers concluded that this single sentence—and associated behaviours resulting from it—constituted obstruction of justice because it encouraged destruction of evidence that Quattrone had by then known was being requested by the government (Gasparino 2003).[1] The lesson is clear: a single email message would never have been enough to warrant an interpretation, legal or historical. But presented in context, an eight-word message can be seen as evidence of innocence or guilt. Getting context right is always a challenge for scholars, one made even more difficult when dealing with decontextualised artifacts like massive email corpora.

Electronic mail came to replace most day-to-day forms of organizational correspondence by the late 1990s (Kirsch 2009; Moss 2012). Since then, its importance as both a personal and institutional communication medium

✉ Stephanie Decker
stephanie.decker@bristol.ac.uk

David A. Kirsch
dkirsch@umd.edu

Santhilata Kuppili Venkata
santhilata.venkata@nationalarchives.gov.uk

Adam Nix
a.nix@bham.ac.uk

1 University of Bristol, Bristol, UK

2 Robert H. Smith School of Business, University of Maryland, College Park, USA

3 The National Archives, Richmond, London, UK

4 University of Birmingham, Birmingham, UK

[1] The prosecution was complex and expensive. The first trial ended with a hung jury. The result of the second trial, in which Quattrone was found guilty, was overturned on appeal, following which a settlement was reached out of court (Smith 2006).

Published online: 31 December 2021

has continued despite the rise of other technological solutions, as the debates during the 2016 US presidential election showed. However, because privacy is hard to protect in email, it is predominantly held in dark archives, which effectively preserve historical material but limit access to it (Jaillant 2019; Prom et al. 2018). While these privacy issues are undoubtedly key to mediating research access to (currently closed) email collections, another is how users will actually engage with such material once these access issues have been navigated.

Even those email archives already available (Enron Email Corpus, W3C, ePADD) are often opaque and difficult to search effectively due to the networked nature of email. Specifically, email is a hybrid artifact: email IS and email ARE, as our anecdote up front illustrates. Using keyword searches, performing word frequency counts, network analysis (Aven 2015) and tracking the timing and sequence of email interactions (Byun and Kirsch 2020), each engages with just one aspect of the networked resource that is email. In these respects, email is an inherently richer historical source than pre-digital correspondence, as it does not just contain information as content, but also represents the flow of information between individuals or in organizations as context (Yates 1993; Yates and Orlikowski 1992). Yet this means that for email to become a historical source that can be interpreted by future users, both individual and network aspects of email need to be maintained.

Archival users vary in their familiarity with technological solutions, with Talboom and Underdown (2019) categorising users of digital collection into three broad types, arrayed in terms of their level of digital engagement: (1) "readers", who want to access a digital source like a traditional paper source; (2) "digitally curious", who want to search large databases to identify items of importance for more in-depth study; and finally, (3) "data users", who want to perform computational analysis over entire collections. And while there are some solutions available for groups 1 and 3 (reproduction of digital items where these can be released and access to a large-scale database for computational, non-consumptive use, respectively), there are currently no tools or access options for users in group 2. However, it is likely that many future users will fall into this category, replacing or converting the "readers", while the "data users" are likely to move on to ever bigger and more diverse data sets that continue to push the boundaries of "big data" computational analysis (see Jacobs and Watts 2021).

This shift in terms of user requirement is yet to be fully appreciated, as archivists are focused on the issues surrounding digital preservation, which has required significant changes in workflow. Yet as Talboom and Underdown (2019) point out, access is what is being preserved, and understanding the possibilities for enabling discovery of digital resources in the future is an important issue to consider in conjunction with preservation. AI offers potential solutions here, as they can enable archival users to connect the context around people and events during a certain time period with the content of the emails. We define AI tools as "computer systems that can perform tasks normally requiring human intelligence" (Oxford Reference 2021). This can be achieved by creating a tool based on a knowledge extraction layer over the email network using artificial intelligence for natural language processing (Devlin et al. 2018), thus combining several different analytical approaches to achieve more relevant search results. Such tools would improve discovery for archivists and archival users in terms of identifying relevant content of large email archives for further investigation.

Our paper is, to the best of our knowledge, one of the first to consider the issue of content discovery facilitated by AI tools for digital archives, and we aim to shift the current concern with digital preservation towards a greater recognition of the future challenges of context-sensitive discovery in large-scale digital collections. Our argument proceeds as follows: we first review some of the literature on the role of digital archives in scholarship, including some of the key reports by and for archivists engaging with the challenge of integrating digital resources into existing collections. In the main section, we first describe the email collection that we are working on as part of a larger international, research-council funded research project, and then describe the AI tools that we developed to make the email collection more accessible. We close with a discussion of the main issues and concerns that our work so far has raised and highlight the importance of ongoing work in this area.

2 From preserving emails to using email archives

For more than 20 years, email has represented a predominant communication medium for both individuals and organizations, and is widely accepted to be among the most important of post-analogue technologies (Kirsch 2009; Prom 2011; Rosenzweig 2008). While its cultural and historical significance is undisputed, those responsible for preserving email have to contend with emails' complexity, which primarily comes from its scale and networked nature as a technology and form of correspondence (Prom et al. 2018). Indeed, traditional archival processes have been materially challenged by the requirements of email, particularly in relation to privacy and sensitivity (Shilton et al. 2017). As a form of communication, it is both highly personal—even within organizational contexts—and surprisingly messy. While this is also true of some paper-based forms of correspondence that email is so often likened to, the sheer scale

of traffic makes traditional methods of appraisal unfeasible for just one email account, let alone a corpus of many:

> "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" (Prom et al. 2018).

This perfect storm of scale, entropy, and potential sensitivity has made archives and their potential donors wary of liabilities associated with email archives, a contributing factor to the rarity of access seen today. As we highlight in this review, the complexity of managing email collections, along with their limited availability for research, has meant that research engagement with email has been largely subordinated to issues of preservation and processing (BDAWG 2019).

For many of those interested in using email, its scale and complexity are also going to present challenges to the predominantly analogue assumptions of their practice (Fellman and Popp 2013; Jaillant 2019). For instance, like traditional archival methods, normal historical approaches are framed around the use of paper-based collections, a characteristic suggestive of an access that is physical as well as one based on close reading and finding aids (Howell and Prevenier 2001). Such access is largely sub-optimal for email, which can be searched computationally, made available remotely, and is generally unsuited to close reading alone. It is therefore key to understanding how researchers will engage with email collections and develop appropriate forms of access to support this. To this end, Jaillant (2019) suggests three strategies for improving this engagement. First, as contemporary archives (digital or not) will always carry data protection risks, archives need to trust scholars and their ethical codes rather than trying to remove everything that is sensitive. Second, technical infrastructure need not be perfect, and some access is nearly always better than waiting for the perfect tool. Third, rather than being passive users, researchers should be actively involved and empowered to shape the nature of access. In short, the problem of email will require exploratory collaboration both up-stream and down-stream of accessibility.

While historians are predominant among the stakeholders of born-digital archives, they are yet to engage significantly with either the post-analogue past, or how it will be researched and written (Milligan 2019). However, there are several active areas within the discipline that are engaging with these issues, and many represent opportunities for collaboration. Among these, "digital history" was the first to deal with the methodological and practice implications of digital, emerging via the broader personal computing and internet revolution of the late 90s (Dougherty and Nawrotzki 2013). Interest here was primarily in how digital technologies could be used in research and engagement around traditional, pre-digital historical contexts. However, it was also widely acknowledged by digital historians that, in addition to the tools they made available, digital technologies were also going to materially change how historians viewed the past (Ayers 2001; Cohen et al. 2008; Duranti 2001; Rosenzweig 2008). The related area of web history has developed these ideas further, focusing exclusively on born-digital sources and their potential impact on histories of the digital era (Brügger 2012; Brügger and Milligan 2018; Milligan 2019). Importantly, these potential user communities have also acknowledged the need for historical users to play a more active part in the archival process.

Elsewhere, projects focused specifically on engaging users have already shown the potential value in terms of understanding access to born-digital collections. In 2018, the Wellcome project reported on an initiative between historians and archivists, which sought to explore the impact born-digital archives will have on the historical methods (Sloyan et al. 2018). Importantly, it highlighted that many researchers have limited experience of using born-digital sources and can only speculate about how they would like to access and use them. What was also apparent was that, to contextualise what they find, researchers need information about the digital structure of born-digital records as well as other contextual triangulation points. Similarly, in 2016 the British Library (UK) ran a pilot project that evaluated reading room access to several born-digital collections, testing both their model for processing born-digital material and their provision of access (Jaillant 2019). Insight from coordinated efforts like these can also be combined with the growing body of researcher-led reporting of their experience with born-digital access (Milligan 2019; Nix and Decker 2021).

While debates still revolve around preservation and appraisal, Talboom and Underdown (2019) highlight that, from a more long-term perspective, any type of preservation should ultimately be focused on access. Moreover, for access to be meaningful, it needs to account for the nature of different user groups and their diverse needs. Within digital preservation for cultural objects, an appreciation of different types of user requirements is well established, as well as the need to understand factors that will affect individuals' information-seeking behaviour (Beaudoin 2012). These might include a user's work role, their intended task, or indeed, their level of skill (Lynch 2002; Talja et al. 1999). Along these lines, Marchionini and Maurer (1995) identify three basic users (formal, informal, and professional) in their work on how digital libraries can create integrated resources in academic settings. Similarly, in proposing an integration of

'use history' (e.g., the history of an object's use) as context, Benoit (2011) classify users into three alternative categories: expert, general, or casual. Here, an 'expert' user will have greater knowledge about the data, while more 'general' users do not have such specialised knowledge and thus pose a broader range of (unusual) questions to satisfy their informational needs.

Talboom and Underdown (2019) provide their own typology of users, suggesting that three different types of researchers interested in digital collections are emerging. They stretch in a continuum from the traditional researcher keen on accessing digital sources as an emulation of the experience of the original custodian or reader of a document, to the technologically savvy researcher seeking to access bulk downloads of "big data". The latter type of usage is often referred to as "non-consumptive", meaning the researcher does not read or otherwise process the content of the archive, but rather its meta-data and other structural or semantic features (Jett et al. 2016). In terms of email archives, this may be the network pattern (Aven 2015) or the timing of responses (Byun and Kirsch 2020), which takes a quantitative analytical angle on the large-scale characteristics of the data set that individuals sending and receiving those emails may not have been wholly or partially aware of.

In between those two extreme-type users Talboom and Underdown propose a third group—the "digitally curious"—who would want to use the opportunities of better search and make lateral connections that digital sources afford historians, without wanting to sacrifice the content of such historical documents (see Nicholas and Clark 2018). Yet much like the archivists seeking to appraise and manage such collections, the digitally curious will have to contend with both the abundance and the lack of easily comprehensible organization of such a digital archive. Whitelaw (2015) describes such users as "information flaneurs", and details the principles and development of generous search interfaces in response.

Our own communities of business history (Decker et al. 2015) and historical organisation studies (Maclean et al. 2016) offer good examples of this third user group. As the post-analogue past becomes more historically relevant, there is a growing awareness of the link between digital technologies, and the importance of continued preservation and use of organisational records (Nix and Decker 2021). Indeed, when we ran a roundtable on digital sources at the Business History Conference, the diversity of approaches already being used in relation to organisations was readily apparent.[2] Among the speakers, Gavin Benke, showed how he used relatively traditional methods to analyse Enron's emails, manually iterating between computational search and close reading as his analysis developed (Benke 2018). By contrast, Tim Hannigan's approach used what he terms as rendering, whereby textual data are computationally "prepared" for analysis, with the researcher making analytical choices that determine how and what topics are identified (Hannigan et al. 2019). As with some of our own research using born-digital sources, these approaches evolved specifically to account for the content within digital documents and its relationship to broader structural and technical characteristics (i.e., its context).

As we start to think more about access and uses of email archives, we believe the different ways in which users wish to engage with email data, their technical proficiency, and their knowledge of a collection's content, will become increasing key. It is upon this premise that the prototype we describe in the next section is based, combining interests of context and content into a purpose-built tool for users, such as ourselves to access and search organisational emails.

3 Approaching of email archives

Our project used an email archive already licensed for computational uses through the Linguistic Data Consortium of the University of Pennsylvania, known as *AvocadoIT*. We introduce this collection in the next section and highlight its relevance to more qualitative content-based enquiry. Following that we describe our approach to developing an AI tool that combines content with context, and how this could significantly enhance email archive discovery going forward.

3.1 The email archive: *AvocadoIT*

The organizational email collection that we used for our project was licensed for scientific use by the Linguistic Data Consortium (LDC) of the University of Pennsylvania in 2015 (Oard et al. 2015).[3] The collection is known by the pseudonym AvocadoIT. The company was a venture-backed technology start-up operating in Silicon Valley during the dot.com boom and bust. Data we are using were extracted from a binder labelled "Email Backup" containing c. 20 CDs. The binder was in a box of other technology-related materials recovered in the process of "unwinding" the firm following its failure in 2003. These materials were stored and subsequently discovered by one of the authors. This "backup" predates the ultimate failure of the firm by several months. As a result, many important events occurred outside the timeframe of the recovered email, limiting its ability to shed light on critical questions of scholarly interest surrounding the final months of the firm's existence.

[2] Decker et al. (2021).

[3] Since being made available, it has been licensed.

Nevertheless, the scope of the collection is sufficiently broad and the coverage sufficiently complete that it can be usefully exploited as a historical resource.

The 278 email files (PSTs) on the CDs were subsequently processed with funding from the U.S. National Science Foundation (NSF) under grant #1065250 ("Development and Evaluation of Search Technology for Discovery of Evidence in Civil Litigation").[4] A risk assessment determined that several categories of information needed to be either redacted or removed. Redacted data included personal data, and removed material included image files that contained no machine-readable text as well as emails with the venture capital investors on the board of the start-up. Given that the stated purpose of the collection—the reason NSF had funded the project—was to help advance the field of e-discovery, the decision to withhold emails between the company and its venture investors was a reasonable and necessary accommodation to the data supplier.

3.2 Introducing AI for email contextualisation

Currently available tools[5] for email archiving follow an ingest-and-preserve approach and support information managers if they need to revisit email for legal, litigation and recovery purposes later. Each preserved email provides us with metadata, such as sender, recipient, date-time sent, cc, bcc, subject, attachments, in addition to other technical details. The existing email archiving tools, such as ePADD, etc. migrate and normalise emails to standard-based targets using metadata. The existing tools offer advanced facilities such as:

- turning unstructured into structured data with the help of metadata,
- allowing users to extract conversation utilities,
- assisting with the identification of email file formats,
- parsing messages to identify senders and recipients using text analytics (named-entity-recognition),
- grouping of entities and attachments.

However, these tools are not sufficient to retrieve contextual information for users who want to conduct research on the content, considering the entire email corpus as a single repository.

Using AI tools for the identification and presentation of contextual information for email archives can potentially facilitate both preservation and access. As noted above, email has network properties, and individual emails need to be understood as part of conversational threads. However, it creates two challenges: first, assuming the would-be reader does not choose to read emails selected at random, the emails need to be presented to the reader in a ranked order of relevance. The relevance judgments must incorporate contextual knowledge, but some aspects must be provided by the user. For instance, is the reader interested in the "next phase of our marketing platform" or "our new technology plan"? In either case, the search environment must return a relevance judgment to the user based upon relatively loosely structured queries, as readers will likely be equally unsure in early stages of their exploratory searching what they are looking for.

Second, by treating each email message as a single artifact, existing approaches to email archiving, such as ePADD and DarcMail, tend to decontextualise the email corpus. Even where the "thread" structure of messages is acknowledged and navigation within the thread is possible, further technical issues such as the need for deduplication or named entity resolution may compound the challenge facing the human user seeking to "read" an email exchange in the same way it would have been read by the original participants. For example, who are the email correspondents and the individuals' names in an email or thread? In general, reading email poses evident challenges: for instance, should the reader read the thread from oldest to newest, in the manner that the exchange occurred? If email is a "conversation," then surely this is the proper way to experience the exchanges that took place. No one wants to hear a conversation occur in reverse order. However, in dealing with historical email threads, as in our daily organizational lives, we may wish to move backward in time through the thread, starting with the most recent messages and then reading the earlier messages already cued to look for relevant content. It is not at all obvious how email should be read. Structural analyses have dodged this issue by ignoring email content and focusing only on metadata (i.e., "To", "From", etc.).

In the next section, we describe the implementation of two models using state-of-the-art techniques of Natural Language Processing (NLP) and attention (Vaswani et al. 2017)-based models. The need for advanced techniques can be understood by thinking about the two sample queries above: a simple bag-of-words-based approach might find that "platform" is associated with technology and "plan" with marketing. A context-sensitive approach should return more relevant email threads, such as related and synonymously used terms (e.g., e-business trends, smart market trends, online market exploration). The models developed return sets of emails relevant to the search query, and we evaluate their suitability based upon our subjective satisfaction as prospective users.

[4] Collected almost 20 years ago, the collection is quite small compared to contemporary organizational email collections.

[5] Eas, ePADD, Darc Mail, Apache Tika, FTK.

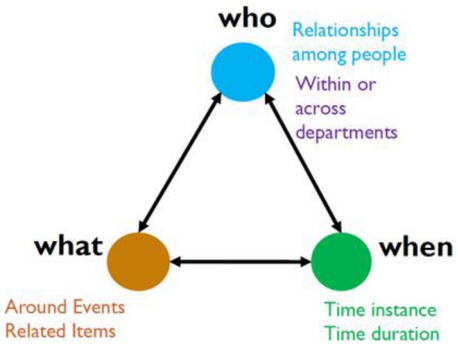

Fig. 1 Context for email archives

3.2.1 Defining context for email archives

Context can have different meanings regarding digital archives. We seek to combine a range of contextual aspects in the development of our discovery tool. In information retrieval (IR), seeking the context of a digital object depends on two processes: digital curation and reuse of the object. While curation focuses on metadata for technical aspects in digital preservation, reuse concentrates on extracting the meaning through metadata (Faniel and Yakel 2011). Essentially, for a digital object, the context is all known properties associated with it and all operations that have been carried out on it (Brocks et al. 2009). Talja et al. (1999) suggest that the context represents some kind of background to understand the patterns of behaviour and social and cultural meanings and values.

Mayer and Rauber (2009) introduced a semi-automatic approach to determine the creation and usage context of digital objects. Various aspects of context in different dimensions such as the time of the object creation and modification, the object type, people involved and content type across different sub-categories, such as the topic, genre, and acronyms used. They believe that digital preservation must provide the tools and techniques to support users' analysis and interpretive processes for digital objects where context is not obvious to the users. Their use case with email mailboxes and attachments provides one basis for our definition of context for emails.

We depart from existing approaches to context in our project, and instead draw on historical practice of source analysis. This is based on a series of critical questions (Howell and Prevenier 2001; Kipping et al. 2014), such as who has written a text and who is it addressed to (or in this case also sent to in BCC), when was the source created (relative to the events it reports on) and what is the content and what was happening at the time that is referenced in the source. We draw on this common methodological practice in our definition of context for email archives as the interdependence of three "entities": who, what and when (see Fig. 1). We define entities as uniquely identifiable objects or things (such as persons, organizations, and places), characterised by their types, attributes, and relationships to other entities. The entity who represents the people who are connected with a particular event (what) during the time instance (when).

The entity who refers to the people who are either senders, recipients or persons referenced within the email contents. The relationships among people in the set of senders, list of receivers, receivers in the "cc" list and "bcc" list can be generated and maintained as a separate registry, thus providing additional information for analysis. The names thus extracted computationally (using, for example, named-entity-recognition) are crucial to the ability of the user of the collection to navigate through the network graph. The entity is what refers to an event or process. The events and processes are useful to connect to other external events, and thus may eventually help to linking beyond a specific dataset. We obtain specific events from the text of the body and attachments using methods of natural language processing. The entity when is the date and time of the event and refers to the time instance of the email sent or received. We demonstrate how context would facilitate a sample user query thus:

> User query: "Get emails from E-business trends between 2001 and 2002"

The above query searches for the "event" of "E-business trends", between 2001 and 2002. The user expects the result set to contain all relevant email threads among multiple senders and receivers around the topic of E-business trends. The user might need to see emails as varied as:

1. emails that refer to the meaning of E-business,
2. examine the set of senders and recipients of these emails to understand who are the individuals involved in the conversation and dates referred in the query, and
3. what other contemporary events are discussed along the E-business.

In summary, our three questions of who, what and when define the context for emails using structured data and extracted information as features. This definition of context represents latent information across one or more conversational threads over a period of time, and goes beyond using metadata alone, which cannot retrieve contextual information. For this, we require an AI tool that comes closer to "reading" the contents of the email.

3.2.2 Email Contextualisation discovery tool

The above-defined context represents a compact set of attributes of email documents in information retrieval terms. It aids to retain the network properties of the email corpus,

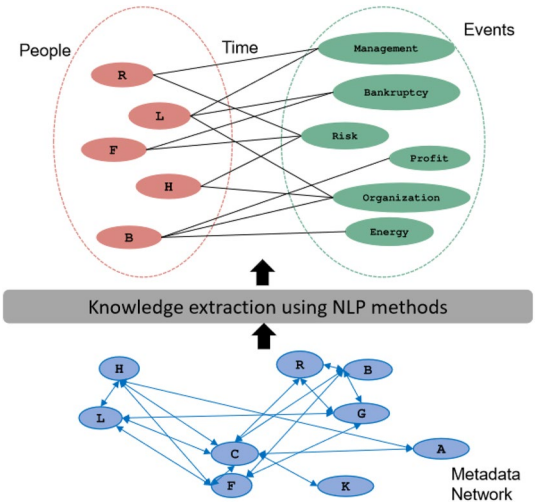

Fig. 2 Knowledge abstraction from the email network for context extraction

while adding the extra information about events (what part of the context) to the graph structure. The knowledge extracted links to what was happening at the time of the email and how close it is in time to the events mentioned in a given email. The knowledge abstraction from the contents of the email archive and the construction of the Email Contextualisation Discovery Tool (EMCODIST) are detailed in this section.

We refer to knowledge abstraction as a process of extracting information about various events, people involved, incidents and interesting developments happening within the organisation. Usually, emails are represented as graph structure with metadata properties (Chapanond et al. 2005; Fu et al. 2007; Laclav'ik et al. 2012). The detailed process of knowledge abstraction is shown in Fig. 2. The nodes of the metadata network are the properties of interest connected to other metadata properties, such as sender/recipients, subject, sent date, etc. The edges represent the strength of relationship by graph metrics, such as degree distribution, diameter, average distance, and compactness between nodes.

We then implement a knowledge abstraction layer over the network graph to extract the contextual information and connection between the people and events on a timeline. Here, we follow a three-step process to abstract knowledge from the email's body content and attachments:

1. *Extraction of metadata properties* Each email is assigned a unique identifier. While the unique identifier acts as a primary key to locate a particular email, other metadata provide us the person-to-person network relations, and person-to-email affinity.
2. *Identification of topics* We used Natural Language Processing (NLP) to identify names and events within the subject, body content and attachments. Given that labelling each of the events or names would be unmanageable across the collection, we took an unsupervised approach to produce a set of high level of topics. The events, names (of the people in the body content) and dates (mentioned in the body content) were grouped to generate a topic repository of 50 + topics. (The number 50 is an optimal number chosen in accordance to the available computational power for the prototype.)
3. *Construction of knowledge graph* The topics are connected to the sender/recipient (through the email iden-

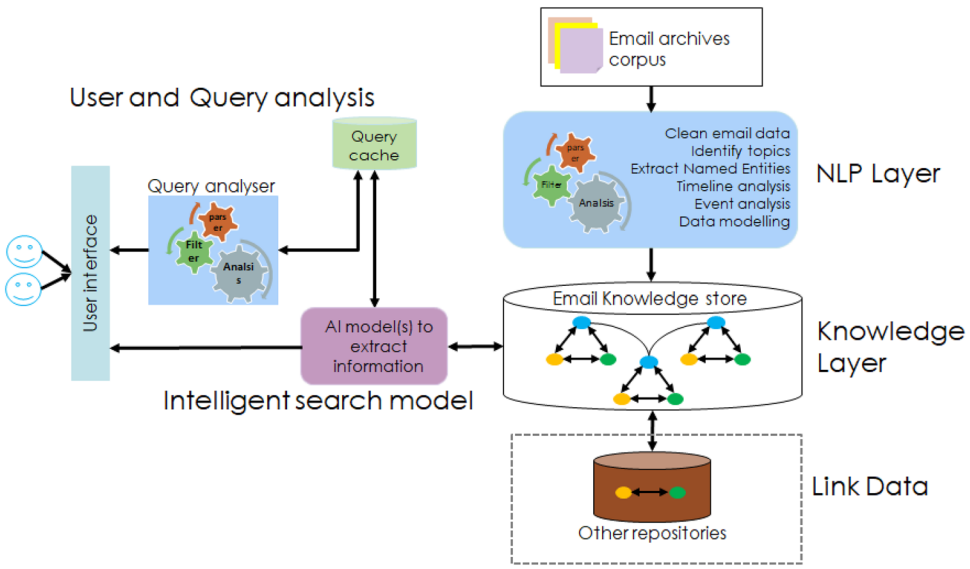

Fig. 3 Overall architectural diagram of the contextualisation tool

tifier) to create a bipartite graph[6] to optimise computational time. The AI models leverage the knowledge graph to find emails within the context of the query.

The overall architectural block diagram and workflow of the discovery tool is shown in Fig. 3. It reflects the user involvement with the machine processes to extract and represents the required contextual information.[7]

User and query analysis is an adaptive system to understand types of users from their query patterns. When a user submits a query, the query analyser parses it to extract names and events to understand the meaning and the context. To leverage a variety of queries and search patterns, the query analyser saves the results of frequently asked queries in a cache[8] to reuse the contextual information.

Data processing and knowledge abstraction The email corpus (AvocadoIT in our case) is passed through a metadata extractor[9] and NLP layers to process the data to use for the knowledge abstraction. An email knowledge graph is constructed as a person-to-event bipartite graph as explained above.

The development of intelligent search models This is the central part of the tool. Based on the user types, we divide their queries into two categories: (1) expert queries and (2) novice queries.

The expert queries are made by a user with some knowledge about the corpus. They are expected to make use of specific keywords and phrases in their queries. The novice queries are made by an inexperienced user to get a feel of the collection. To cater to the needs of both types of users, we have developed two models. Both models identify appropriate email threads and patterns relevant to the query. Our aim is to present both models to users, and to track their choices of model and queries to identify how the discovery tool is being used to improve the efficiency of the search queries over time.

Model 1 is a phrase search model that matches the phrases in the query to the content of the emails. This model is an improvement over the basic keyword search and suitable for expert users. The keywords and phrases from the query are identified with the help of NLP and the model returns all emails and threads that contain the phrase as a single unit.

Model 2 finds similarities between the topics discussed in the corpus and user queries and ranks emails from the query perspective. This model is technically more complex than the first one. It makes use of modern technologies such as an open-source neural network-based technique for NLP, pre-training Bi-directional Encoder Representation from Transformers (BERT)[10] technology[11] to understand the meaning of the words in context to their neighbouring words in an email's content. The model matches the central meaning of the email thread to the meaning of the query.

The discovery tool developed as a proof-of-concept is presented in Fig. 4. The source code jupyter notebooks have been made available on Github.[12]

3.2.3 Evaluation of the models

We conducted an initial evaluation of the email results returned by the models for their contextual similarity to the query submitted. A sample evaluation on a set of ten queries (on *Y*-axis) is presented in Fig. 5. The bars represent the percentage of relevant emails from the entire result set. We have chosen the queries to represent three categories of queries submitted by different user types.

It is observed that model 1 performed well for queries, such as "technical support", "weekly updates", "E-business", which have a good number of appropriate results with the help of the extra input provided by the user. The extra information includes the (1) choosing of an appropriate group of mails to search the result from and (2) a probable start date and end date. These queries represent the adaptability of the tool with which it can utilise a user's knowledge to add the context. The keywords used are specific to the email corpus and the context in which they are used. This requires the user to be familiar with the language and terminology used in the corpus, in this case the organization's "slang" and abbreviations, as well as practices (e.g., "weekly updates" emails sent by company executives). Because AvocadoIT is a technology company, many emails refer to notifications of technical app failures and conversations about wireless technological solutions developed. Knowing this context will allow users to craft better queries when using model 1.

[6] In the mathematical field of graph theory, a bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets.

[7] Assumptions and limitations:

1. The language of conversation in the emails was restricted to English only.
2. Since the tool was developed to search contextual information, we have omitted auto-generated emails from Tech help desk and servers from the data corpus.
3. The models developed are not supporting special characters at the moment.
4. Models are tested on the base model of BERT to keep the computation to a minimum in this version of proof-of-concept.

[8] A memory location set for quick reference. Mainly used for frequent queries.

[9] A python package email.parser facility is used.

[10] BERT is used for computational processing of natural language, see: https://blog.google/products/search/search-language-understanding-bert/.

[11] https://github.com/google-research/bert.

[12] https://github.com/Contextualising-Email-Archives/discovery-tool.

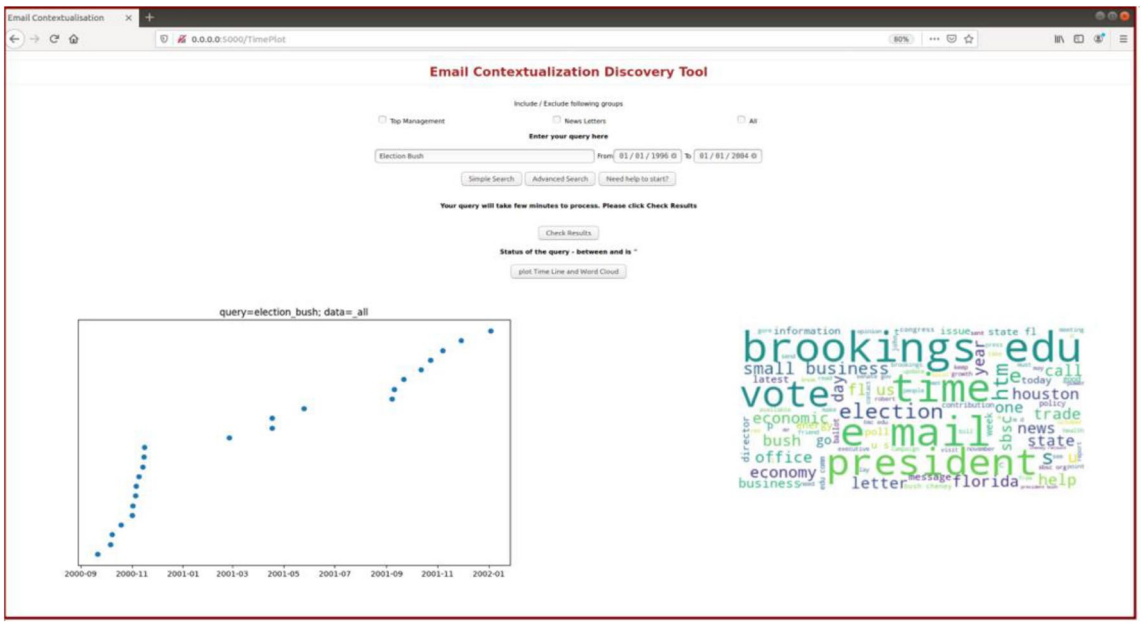

Fig. 4 The email contextualisation discovery tool (EMCODIST)

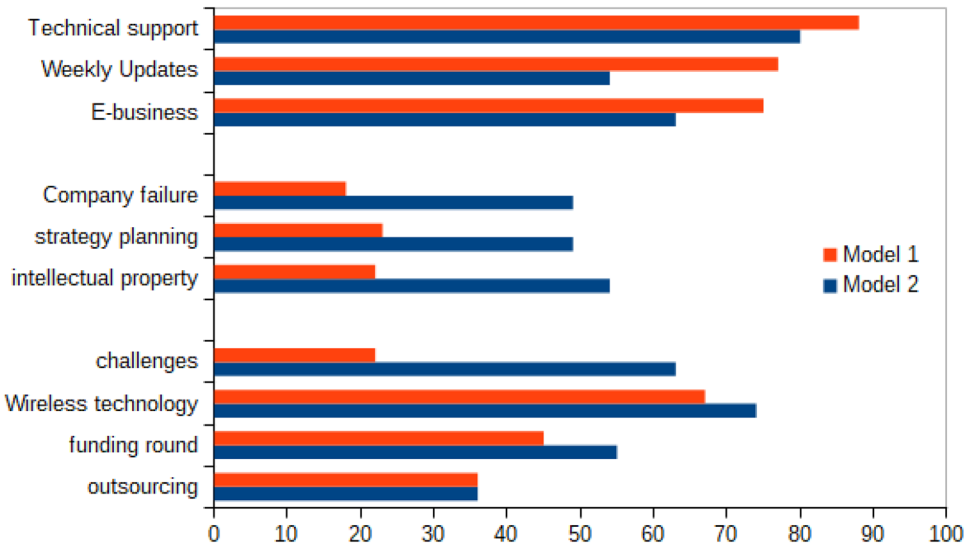

Fig. 5 Fraction of emails contextually similar to the sample queries presented

For more generic queries, such as "company failure", "strategy planning", and "intellectual property", model 2 returned a larger number of emails with relevant meaning. The conversations in the email threads never mentioned these specific keywords, but their general theme of email is around the keywords. We expected this because BERT embeddings are designed to facilitate such “interpretation”.

The third category of queries like "challenges" has multiple meanings and hence the result set has many topics belonging to variety of challenges. For such queries, a user may want to run model 2 first to get an idea of the corpus to narrow down the search with model 1 later. We summarise the characteristics of both models in Table 1.

4 Discussion

4.1 Configuring the model/user

Our approach has introduced two distinct AI-based methods to support user-driven exploration in email corpora. These two models—one based upon simple, phrase-based

Table 1 A summary of both models

Concept	Model 1	Model 2
Model technique	Phrase matching by NLP techniques	BERT embeddings for document Classification and development of knowledge graph
Computational time	A very low response time (almost instantaneous)	Response time depends on the size of data corpus (computational time needed for document distance calculation)
Memory requirements	Volume of data corpus: specific phrases (upto 800 MB for AvocadoIT)	Volume of data corpus + word embeddings generated by BERT(up to 2 GB for AvocadoIT)
Processing requirements	Can work with processors similar to intel core i5 models	Needs a GPU processor to run BERT models
Model usefulness	Works well with small volumes of data	Can handle medium to large data volumes
Fairness and 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
Accountability and error corrections	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)
Working with higher models of BERT	Not applicable	Can provide more accurate results with higher versions
Interpretation of the model result	Results obtained are the set of all emails that contain the keywords presented in the query. The mails are ordered by the sent time	Results are a set of emails rank ordered by the similarity of its contents to the query
User type	Knowledgeable users who have some idea about the contents of the corpus	Suggested for novice users without information about the corpus

matching and the other upon more complex, BERT embeddings—generate systematically different types of results. We have hypothesised that each tool is better suited for distinct classes of users, here envisioned as context knowledgeable (experts) and context unaware (novices). In this, we extend some previous work on user types (Beaudoin 2012; Benoit 2011; Marchionini and Maurer 1995) in that we see these types as dynamic—model 2 supports more explorative queries, whereas model 1 allows expert users familiar with the language and the individual names in an email corpus to conduct more focused searches.

This is reflected in our theorization by different types of queries. For queries that can be formulated using relatively simple, descriptive terminology (such as "weekly updates"), the phrase-based matching tool (model 1) returns more results than the BERT-based one. Conversely, when searching for less precise concepts like "failure" or "strategy," the BERT-based tool (model 2) returns more results. Importantly, we have not pre-processed the collection to establish a truth table that would allow firm conclusions with respect to these results. Rather, our findings suggest that specific types of users will be better served by different AI-based search methods and that users themselves may be incapable of knowing ex ante which AI-based search method would best suit their needs. Importantly, using different AI-based approaches in combination allows users to become more knowledgeable about a collection through their explorations. Such tools have the potential to become "finding aids" in their own right, and may be able to provide guidance to complex collections (Sloyan et al. 2018).

The successful application of AI models for contextualisation of email depends upon decisions made about the initial rendering of the corpus. In the case we have considered, the AvocadoIT collection was produced for use as a test collection in the information retrieval community. Considerable effort had already been expended to make the collection usable for that community of researchers. Fortunately, and notwithstanding several important caveats that we have noted, that rendering did not limit the usefulness of the collection for our purposes when trying to apply AI tools for discovery in service of business history, a subset of humanities research. However, we should not assume that this relationship will necessarily hold for future collections. Some proposed schemas for conducting non-consumptive research in email corpora might—by design—limit the fruitfulness of a given collection for downstream use in the humanities. Our example underscores the challenge of imagining, identifying, and enabling the fullest possible range of interests that users bring to a collection.

4.2 Individual vs. organizational email

Email corpora vary. Some map easily to pre-digital archival practices. For instance, an email corpus belonging to a single individual can be treated as a collection of correspondence (Schneider et al. 2019). Respondents can be identified, rights requested and granted or denied, and access managed according to updated, but otherwise known practices. Many humanities collections fall into this category. New insights are likely to be based upon content analysis, or highly specialised interpretation of specific exchanges. The email corpora on which we are working, however, do not share these features, as they were generated by organizations. Organisations organise; they bring individuals together to pursue common purposes that are beyond the capability of individuals. Even though many organizational emails are sent by individuals, others are sent by individuals acting organizational roles, as when an employee monitors and responds to email sent to an "info@company.com" mailbox. Still others are generated automatically, in response to online orders and as digests and summaries of other, non-email-related actions. In organizational settings, individual conversations, like those found in personal email archives, are embedded in multiple layers of context. As a result, surfacing meaning from complex corpora such as organizational email collections raise problems that are not easily characterised, let alone solved.

First, organizational email corpora are too large to be read in their entirety by a single researcher or research team (Prom et al. 2018). Whereas a competent scholar could probably make their way through the correspondence of a single email collection (perhaps with a little bit of keyword-based pruning), the AvocadoIT collection includes close to 1 million messages. It is not reasonable to imagine that even a dedicated scholar could "read" the entire corpus. Nor would a scholar want to read all million messages. The challenge becomes figuring out which subset of messages the scholar wants to read, based upon the nature of the question that is being asked. Studying organizational email necessarily places the scholar in a context of data overabundance (Czarniawska and Löfgren 2013) where search and selection are paramount for sensemaking and interpretation.

Second, organizations are studied for many reasons and by scholars from multiple analytic traditions. Traditional archival appraisal and accession practices always foreclose some questions even as they privilege others. It is impossible to anticipate multiple questions that a scholar might bring to a collection. Our discussion of two different types of user highlights that each email collection is, first, discovered by someone who is new to it and needs to explore it to ask better questions, and second, later searched by users who

have become more familiar with its content and context. The potential users of organizational email are clearly going to be diverse in terms of their knowledge and the questions they are going to ask of the material.

Third, an essential feature of an organizational email collection is that it includes many messages that are not relevant to the organization itself, except in so much as the email shows that the organization existed in a particular environment. These newsletters or those external events were being shared and reported about in this setting at this time. A scholar interested in life in America in the year 2000, for instance, might be very interested in the conversational language employees used when interacting, the personal news they chose to share and converse about, or the products advertised in newsletters that were circulating within AvocadoIT. However, that same scholar might be relatively indifferent to the particular design decisions that were made about the firm's software architecture or the selection of one marketing partnership over a competing one. For these reasons, we believe that emails are particularly useful, hybrid artifacts for scholars interested in understanding not just business aspects, but also in critically investigating cultural elements of organizational life. We based our discovery prototype on this understanding of a diversity of users, and the assumption that a combination of content and context is the basis on which digitally curious users (Talboom and Underdown 2019) will explore the usefulness of such collections for their own interest.

4.3 Limitations and future research

To realise the full potential of email archives, we need to pursue a form of access that both reflects the nature of email as a unique born-digital medium, and accounts for the various ways in which users will engage with them. The potential for linking content and context goes well beyond our particular focus within this paper. On the one hand, the structural properties of email provide a basis for inclusion and exclusion, which enables a user to define the context in which they search. Equally, machine-based contextual search can provide context based on semantic associations and an appreciation of words within their linguistic context. It may be that by combining relatively open user-led interfaces with machine-led alternatives, digital archives can provide environments suited to both forms of search, allowing for a translation of more traditional historical norms to email and the integration of more novel opportunities for discovery. Additionally, context can be brought in from the outside, in the same way that historians triangulate their discoveries with traditional archival context. While our work here does not explicitly "bring external context in", we see integration of external sources of context (e.g., organizational charts, press media, or financial data) as a vital next step in enhancing discovery.

Another natural extension to our research on the contextual search processes of users, is how the user's inputs and results are presented. The issues of user interface and experience are paramount to effective access, but such considerations go beyond mere user friendliness, and have material implications for how research will read and search email archives. For instance, issues around the ordering of a thread (backward or forwards in time) or the visibility of emails duplicated in and across accounts are not straightforward, as the best presentation of a corpus will depend on the research question at hand. Although the examples of "generous interfaces" presented by Whitelaw (2015) are more visual than textual, the idea of a generous search that allows lateral browsing of a digital archive could also work for email collections. Related to this, it is also important to understand better how our tool performs in relation to users' actual experiences, particularly in relation to the relevance of results. We would like to conduct extensive user evaluation once we are able to provide the necessary access and see this as the next key step in improving overall performance of our tool in the future.

5 Conclusion

Many discussions of email archives still focus on preservation, but this assumes that access becomes logical next step. Unfortunately, due to the unique networked nature of email, and the attendant issues of private and confidential information, this is not necessarily an easily solved problem. It does not only raise the question of who can access the collection, but also what tools are needed to enable effective searching, whilst supporting contextualization. Within this paper, we have focused on one particular form of contextualisation, that of emails' content and its discoverability via an AI-enhanced search tool. While this has yielded encouraging results and insight, it raises new technical and methodological questions around the multiple pathways of discovery and the nature of context as a source of historical understanding. However, the landscape of digital archival discovery is still emerging, and the approaches future historians will take when using digital sources are still unclear. It may be that once various sources of context are combined, researchers will wish to search email in multiple, interrelated ways, requiring access that is both highly customisable but also accessible to novice or casual users. Moreover, these needs may depend significantly on the type of email corpora under investigations: their size, the number of custodians, and the time period that it covers. Such questions therefore not only require that we appreciate how users will search AvodacoIT,

but also many other collections, whether organisational or otherwise.

Acknowledgements We gratefully acknowledge funding support by the Arts & Humanities Research Council (UK) and National Endowment for the Humanities (USA) as part of the US-UK Partnership Development Grants, Grant AH/T013060/1.

References

Aven BL (2015) The paradox of corrupt networks: an analysis of organizational crime at Enron. Organ Sci 26(4):980–996

Ayers EL (2001) The Pasts and futures of digital history. Hist News 56(4):5

BDAWG (2019) Final report of the born digital archives working group email archiving task force. Yale, CT

Beaudoin J (2012) Context and its role in the digital preservation of cultural objects. D-Lib Mag 18

Benke G (2018) Risk and ruin: enron and the culture of American Capitalism. University of Pennsylvania Press, Philadelphia

Benoit G (2011) Integrating use history as a context for dynamically updated metadata. J Libr Metadata 11(3–4):129–154

Brocks H, Kranstedt A, Jäschke G, Hemmje M (2009) Modeling Context for Digital Preservation. In: Szczerbicki E and Nguyen NT (eds) Smart information and knowledge management: advances, challenges, and critical issues. Berlin and Heidelberg: Springer, pp 197–226. https://doi.org/10.1007/978-3-642-04584-4_9

Brügger N (2012) When the present web is later the past: web historiography, digital history, and internet studies. Hist Soc Res 37(4):102–117

Brügger N, Milligan I (2018) The SAGE handbook of web history. SAGE Publications Limited, London

Byun H, Kirsch DA (2021) The morning inbox problem: email reply priorities. Acad Manag Discov 7(2):180–202

Chapanond A, Krishnamoorthy MS, Yener B (2005) Graph theoretic and spectral analysis of Enron email data. Comput Math Organ Theory 11(3):265–281. https://doi.org/10.1007/s10588-005-5381-4

Cohen D, Frisch M, Gallagher P, Mintz S, Sword K, Taylor AM, Thomas W, Turkel W (2008) Interchange: the promise of digital history. J Am Hist 95(2):452–491

Czarniawska B, Orvar L (2013) Coping with excess: how organizations, communities and individuals manage overflows. Edward Elgar, Cheltenham

Decker S, Kipping M, Daniel Wadhwani R (2015) New business histories! Plurality in business history research methods. Bus Hist 57(1):30–40

Decker S, Kirsch D, Kuppili Venkata S, Nix A, Benke G, Ogden J, Hannigan T, Oard D (2021) Making sense of digital sources. Roundtable. Business History Conference. Virtual. https://thebhc.org/meeting-program/19730

Devlin J, Ming-Wei C, Kenton L, Kristina T (2018) "{BERT:} Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.0

Dougherty J, Nawrotzki K (2013) Writing history in the digital age. University of Michigan Press, Ann Arbor

Duranti L (2001) The impact of digital technology on archival science. Arch Sci 1(1):39–55

Faniel I, Yakel E (2011) Significant properties as contextual metadata. J Libr Metadata 11:155–165

Fellman S, Andrew P (2013) Lost in the archive: the business historian in distress. In: Czarniawska B, Löfgren O (eds) Coping with excess. How organizations, communities and individuals manage overflows. Edward Elgar, Cheltenham

Fu X, Hong S, Nikolov NS, Shen X, Wu Y, Xuk K (2007) Visualization and analysis of email networks. In: 2007 6th international Asia-Pacific symposium on visualization. pp 1–8. https://doi.org/10.1109/APVIS.2007.329302

Gasparino C (2003) How a string of e-mail came to haunt CSFB, Star Banker WSJ. Wall Street Journal, February 28

Hannigan TR, Haans RFJ, Vakili K, Hovig T, Glaser VL, Wang M, Kaplan S, Jenning PD (2019) Topic modeling in management research: rendering new theory from textual data. Acad Manag Ann 13(2):586–632. https://doi.org/10.5465/annals.2017.0099

Howell M, Prevenier W (2001) From reliable sources: an introduction to historical methods. Cornell University Press, Ithaca

Jacobs AZ, Watts DJ (2021) A large-scale comparative study of informal social networks in firms. Manag Sci 67(9):5489–5509

Jaillant L (2019) After the digital revolution: working with emails and born-digital records in literary and publishers' archives. Arch Manuscr 47(3):285–304

Jett J, Cole TW, Maden C, Downie J (2016) The HathiTrust Research Center workset ontology: a descriptive framework for non-consumptive research collections. J Open Humanit Data 2:e1. https://doi.org/10.5334/johd.3

Kipping M, Wadhwani RD, Bucheli M (2014) Analyzing and interpreting historical sources: a basic methodology. In: Bucheli M, Wadhwani RD (eds) Organizations in time. Oxford University Press, Oxford, pp 305–330

Kirsch DA (2009) The record of business and the future of business history: establishing a public interest in private business records. Libr Trends 57(3):352–370

Laclav'ik M, Dlugolinský Š, Šeleng M, Ciglan M, Hluchý L (2012) Emails as graph: relation discovery in email archive. In: Proceedings of the 21st international conference on world wide web, WWW '12 companion. Association for Computing Machinery, New York, pp 841–846

Lynch C (2002) Digital collections, digital libraries and the digitization of cultural heritage information. First Monday 7(5). https://doi.org/10.5210/fm.v7i5.949

Maclean M, Harvey C, Clegg SR (2016) Conceptualizing historical organization studies. Acad Manag Rev 41(4):609–632

Marchionini G, Maurer H (1995) The roles of digital libraries in teaching and learning. Commun ACM 38(4):67–75

Mayer R, Rauber A (2009) establishing context of digital objects' creation, content and usage

Milligan I (2019) History in the age of abundance? How the web is transforming historical research. McGill & Queens University Press, London

Moss M (2012) Where have all the files gone? Lost in action points every one? J Contemp Hist 47(4):860–875

Nicholas D, Clark D (2018) Finding stuff. In: Endicott-Popovsky B, Moss M (eds) Is Digital Different? Facet, pp 19–34

Nix A, Decker S (2021) Using digital sources: the future of business history? Bus Hist (advance online) 1–23
Oard D, Webber W, Kirsch D, Golitsynskiy S (2015) Avocado research email collection. Linguistic Data Consortium, Philadelphia
Oxford Reference (2021) Artificial intelligence. https://doi.org/10.1093/oi/authority.20110803095426960. Accessed 27 Apr 2021
Prom CJ (2011) Preserving email. Technology Watch Reports (December):57
Prom CJ, Murray K, Baker F, Connelly M, Gogel W (2018) The future of email archives: a report from the task force on technical approaches for email archives. Washington DC
Rosenzweig R (2008) Scarcity or abundance? Preserving the past in a digital era. Am Hist Rev 108(3):735–762
Schneider J, Adams C, DeBauche S, Echols R, McKean C, Moran J, Waugh D (2019) Appraising, processing, and providing access to email in contemporary literary archives. Arch Manuscr 47(3):305–326
Shilton K, Wickner A, Oard DW, Lin J (2017) Protecting sensitive email: archival views on challenges and opportunities. In: PC4DS: the first international workshop on privacy-sensitive collections for digital scholarship, August 2017, Montreal, Quebec, Canada. Montreal CA
Sloyan V, Demissie S, Eveleigh A, Baker J (2018) Overview of a born-digital archives access workshop. London
Smith R (2006) Court decides for quattrone, reversing verdict. Wall Street Journal, March 21
Talboom L, Underdown D (2019) 'Access is what we are preserving': but for whom? digital preservation coalition. DPC Blog. https://www.dpconline.org/blog/access-what-we-are-preserving. Accessed 30 July 2020
Talja S, Keso H, Tarja P (1999) The production of 'Context' in information seeking research: a metatheoretical view. Inf Process Manag 35(6):751–763
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. CoRR abs/1706.0
Whitelaw M (2015) Generous interfaces for digital cultural collections. Digit Humanit Q 9(1):1–16. Available at: http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html. Accessed 22 July 2021
Yates JA (1993) Control through communication: the rise of system in American Management, vol 6. Johns Hopkins University Press, Baltimore
Yates JA, Orlikowski WJ (1992) Genres of organizational communication: a structurational approach to studying communication and media. Acad Manag Rev 17(2):299–326